目 录

绪　言

——语言学是数学和人文科学之间的桥梁

法国数学家阿达玛(J·Hadamard)曾经说过："语言学是数学和人文科学之间的桥梁"。阿达玛不愧是一位有远见卓识的学者，他清楚地看出了语言学在人文科学中是最容易与数学建立联系的。

然而，在科学发展史上，人们是经过了相当长的过程才认识到语言学和数学之间的这种亲密关系的。

传统语言学的目的在于规定正确的读和写的种种规则，这样的语言学有点象法律。历史语言学用谱系树的方法研究语言的亲属关系，明显地受到进化论思想的影响，这样的语言学一如生物学。结构语言学着力于研究语言结构，力图找出语言中各种要素之间的结构规律，这样的语言学则似化学。

语言学和数学都是有着相当长历史的古老学科。语言学历来被看做典型的人文科学，数学则被许多人看做是最重要的自然科学。在学校教育中，语文和数学被认为是两门最基础的学科，成为任何一个受教育者的必修课。它们似乎成了学校教育中的两个极点：一个极点是作为文科代表者的语文，另一个极点是作为理

科代表者的数学。很少有人会想到，这两门表面上如此不同的学科之间还有着深刻的内在联系。

十九世纪中叶，才有人提出用数学方法来研究语言现象的想法。1847年，俄国数学家布里亚柯夫斯基(В.Я.Буляковский)认为可以用概率论进行语法、词源及语言历史比较的研究。1894年，瑞士语言学家索绪尔(De Saussure)指出，"在基本性质方面，语言中的量和量之间的关系可以用数学公式有规律地表达出来"。后来，他在其名著《普通语言学教程》(1916年)中又指出，语言学好比一个几何系统，"它可以归结为一些待证的定理"。1904年，波兰语言学家博杜恩·德·库尔特内(Baudouin de Courtenay)认为，语言学家不仅应该掌握初等数学，而且还有必要掌握高等数学。他表示坚信，语言学将日益接近精密科学，语言学将根据数学的模式，一方面"更多地扩展量的概念"，一方面"将发展新的演绎思想的方法"。1933年，美国语言学家布龙菲尔德(L.Bloomfield)提出了一个著名的论点："数学不过是语言所能达到的最高境界"。

当时，学者们不仅仅只是提出这些颇具新意的想法，还有许多学者用数学方法对语言进行了实际的研究。1851年，英国数学家德·摩根(A.de Morgan)曾把词长作为文章风格的一个特征进行过统计研究。1867年，苏格兰学者坎贝尔(L.Campbell)用统计方法来确定柏拉图著作的执笔时期。1881年，德国学者迪丁贝尔格(W.Dittinberger)进一步用统计方法把柏拉图著作的执笔时期分为前期、中期和后期三个阶段。1887年，美国学者门登荷尔(T.C.Mendenhall)对不同时期的英国文学著作进行过统计分析，特别是研究了莎士比亚的作品。1898年，德国学者凯定(F.W.Kaeding)编制了世界上第一部频度词典《德语频度词典》，用以改进速记的方法。1913年，俄国数学家马尔可夫(А.А.Марков)研究了普希金叙事长诗《欧根·奥涅金》中俄语字母序列的生成问题，提出了马尔可夫随机过程论。1925年，我国教育家陈鹤琴发表了第一部汉字频率统计的著作《语体文应用字汇》。1935年，美国语文

学家齐夫(G.K.Zipf)发表了齐夫定律。同年，加拿大学者贝克(E.Varder Beke)提出了词的分布率的概念，认为词典选词时，应以分布率为主要标准，频度为辅助标准。1944年，英国数学家尤勒(G.U.Yule)发表了《文学词语的统计分析》一书，大规模地使用概率和统计方法来研究语言。

然而，上述的各种用数学方法来研究语言的想法和具体的工作，都没有对当时的语言学研究发生显著的影响。这主要是由当时的社会实践的要求决定的。因为当时的语言学，主要是为语言教学、文献翻译、文学创作和社会历史研究服务的。在这样的实践要求下，语言学没有多大的必要与数学接近。当然，上述各种研究中不乏卓越的工作。例如，马尔可夫在研究俄语字母序列的数学研究中，提出了马尔可夫随机过程论，后来成了一个独立的数学分支，对现代数学的发展产生了深远的影响。语言结构中所蕴藏着的数学规律，成了马尔可夫创造性思想的源泉。可惜的是，马尔可夫这一卓越的成就，在语言学界却鲜为人知。语言学仍然沿着自己传统的道路，孤立于数学之外，迟缓地发展着。

第二次世界大战以来，由于科学技术突飞猛进的发展，科技文献的数量迅速增加，其增长速度十年翻一番。据联合国经济合作与发展组织估计，从1960年到1985年，世界情报量增加了10～16倍。全世界发行的图书总数是：1952年约25万种，1962年近40万种，1972年约56万种，1980年达到70万种。科技文献的这种增长情况被形容为“情报爆炸”。面对浩如烟海的科技文献，研究人员为了取得全面而准确的科技情报，不得不花费大量的人力物力财力来做难以数计的翻译工作和检索工作，犹如大海捞针，严重地影响了科研工作的效率。

1946年第一台电子计算机问世后，人们开始考虑把这些繁重的工作交给计算机去做，这就提出了机器翻译、机器自动做文摘、机器自动检索科技文献等自然语言信息处理的问题。

在用计算机进行自动翻译的时候，必须进行原语词法、句法

和语义的自动分析以及译语句法和词法的自动生成。这就首先要把这些问题用数学的语言加以描述，从而建立语言的数学模型。

在用计算机自动做文摘和检索时，要求把科技文献的信息储存在计算机中，建立数据库。数据库可以按照人们的要求，在其所储存的信息范围内，对人们提出的问题自动地作出回答。在这种数据库中用以存储信息的语言，在内容上应该是严格的、精确的，在形式上应该适于数据库储存形式的要求，这当然也要求用精密的数学方法对自然语言进行描述。

由于自动化技术和计算技术的发展，人们正迅速地解决生产过程自动化问题，用自然语言来进行"人机对话"，让电子计算机理解自然语言，这就要用数学方法来研究句法结构和语义结构的形式化表达方式以及知识的形式表示技术。

目前微型计算机已逐渐普及，它已经在办公室的事务管理中得到了广泛的使用，这就是"办公室自动化"问题。自动化的办公室要用微型计算机来编辑和处理各种书面文件，这就要求对语言文字进行严格的形式化的描述。

另外，通讯技术的发展，要求对负荷信息的语言寻找最佳编码方法，要求提高信道的传输能力，以便在保持意义不变的前提下，最大限度地压缩所传输的文句，在单位时间内传输最多的信息，这就要求对语言的统计特性进行精密的研究。

在上述的各种促使语言学与数学接近的因素中，最为关键的因素是电子计算机的出现。电子计算机是一种信息处理机，而自然语言是信息的最主要的载体，电子计算机的研制和发展离不开自然语言的信息处理，而自然语言的信息处理离不开数学。语言学家必须采用数学思想和数学方法来研究自然语言，才能回答信息化时代对语言学提出的严峻挑战。

早在现代电子数字计算机出现之前，英国数学家图灵(A.M.Turing)就预见到未来的计算机将会对自然语言研究提出新的问题。他指出："我们可以期待，总有一天机器会同人在一切的智

能领域里竞争起来。但是，以哪一点作为这种竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。"① 图灵提出，检验计算机智能高低的最好办法是让机器讲英语，理解英语。而为了做到这一点，采用数学方法和数学思想来研究语言就是势所必然的了。

由于电子计算机的出现和发展，数学渗透到了语言学的许多领域。具体说来，有如下几个方面：

第一，第一台电子计算机刚问世的1946年，英国工程师布斯(A. D. Booth)和美国工程师韦弗(W. Weaver) 在讨论电子计算机的应用范围时，就提出了用电子计算机进行机器翻译的设想。韦弗还发表了关于机器翻译的备忘录，主张进行机器翻译试验。1954年，在美国国际商用机器公司(IBM公司)的支持下，美国乔治敦大学进行了世界上第一次机器翻译试验，同年，美国海军军械试验站用 IBM701 计算机建成了世界上第一个自动情报检索系统。从此，机器翻译和自动情报检索工作蓬勃兴起。在这样的研究工作中，需要进行词的切分，这就要深入地研究构词法，从而促进了形态学的研究。传统的形态学要区分屈折(inflection)与派生 (derivation)，如英语的 amend/amended 是屈折，amend/amendment 是派生。然而，对于计算机来说，这样的区分是不必要的。一个自动形态分析方案可包含一部词干词典以及一套构词的语法规则（既有派生，亦有屈折)。这样，给出词干，机器可以自动地列举出其可能的屈折形式；给出一个屈折变化的词，机器可以把它分析为词干和词缀。对于机器来说，必须区分各种同形现象，如frighten中的-en要与oven中的-en 区别开来，reaped中

① A. M. Turing, "*Can A Machine Think?*", Mind 50, 1950; 亦见*The World of Mathematics* (ed by J. K. Newman), p.2099

的-ed要与reed中的-ed区别开来。另外，还要考虑一些特殊的现象。如：perform，give，go（现在时）—— performed，gave，Went（过去时）。在去掉词缀之后，有时还要把词干作一定的改变，如cities/city，这样，就要把许多表面上所谓不规则的例外情况条理化。在计算机欣欣向荣的五十年代末和六十年代初，学者们曾对俄语、德语这样一些富于屈折变化的语言进行过严格的自动形态分析，编制了相当精细的自动形态分析规则。

在自动形态分析中，数学方法起着重要的作用。例如，学者们采用离散数学中有限自动机理论来设计自动形态分析模型，从而控制单词的形态切分过程。词可以分为前缀、词干、后缀、词尾几个构词成分。有的词只需要一个词干即可构成，如Work；有的词需要词干和后缀两部分才能构成，如Worker（work-是词干，-er是后缀）；有的词需要词干、后缀、词尾三部分才能构成，如workers（work-是词干，-er是后缀，-s是词尾）；有的词还可以带前缀，如incompact(in-是前缀，compact是词干)。如果把一种语言中词的各个构词成分分别编成词典，如词干词典、词缀词典、词尾词典等，在词典中注明各个构词成分的语法信息，然后，设计如下一个有限自动机来控制切分过程，便可实现单词的自动形态分析：

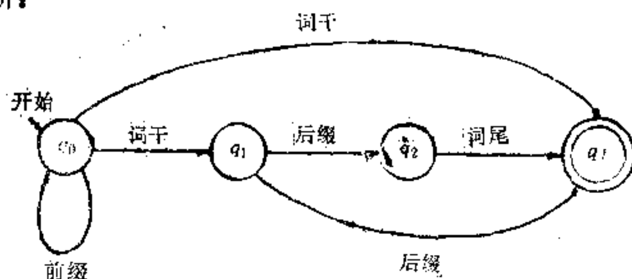


图0.1　切分单词的有限自动机

在图0.1的有限自动机中，q_0，q_1，q_2，q_f 是状态的有限集合，q_0是初始状态，q_f是最后状态，从一个状态过渡到另一个状

态可切分出一个构词成分。例如，从状态q_0到状态q_1以及从状态q_0到状态q_f可切分出词干，从状态q_1到状态q_2以及从状态q_1到状态q_f可切分出后缀，从状态q_2到状态q_3可切分出词尾，从状态q_0返回到状态q_0可切分出前缀。任何词的切分，都是从q_0开始，到q_f结束。例如，词 work 是一词干，其切分状态是 从 q_0到q_f，词worker是由词干和后缀构成的，其切分过程是从q_0到q_1最后到q_f；词workers 是由词干、后缀和词尾构成的，其切分过程是从q_0到q_1到q_2最后到q_f；词incompact是由前缀、词干构成的，其切分过程是从q_0到q_0最后到q_f。

在切分过程中，有限自动机把词典中各构词成分相应的语法信息，记录到输入词中去，这样，当切分结束时，每个输入词都附上了有关的语法信息，为进一步的分析提供了数据。

可见，数学方法的引入有效地解决了单词的自动形态分析问题。计算机的自然语言处理就象催化剂，它促进了数学和语言学的结合。

第二，后来人们发现，机器翻译时不仅要找出两种语言的词汇对应关系，还要进行句法分析，也就是要用句对句翻译来代替词对词翻译，这就促进了自动句法分析的研究。

句法的形式化分析也要求助于数学。苏联数学家库拉金娜(O. C. Кулагина)用集合论方法建立了语言模型，精确地定义了一些语法概念，这一模型成为苏联科学院数学研究所和语言研究所联合研制的法俄机器翻译系统的理论基础。著名的数理逻辑学家巴希勒(Y. Bar-Hillel)提出了范畴语法，建立了一套形式化的句法类型及演算规则，通过有穷步骤，可以判断一个句子是否合乎语法，这些，都大大地推动了传统的句法分析方法向精密化、算法化的方向发展。可见，数学方法的引入给句法的形式化分析带来了生机。

第三，六十年代出现了高级程序语言，使计算机工作者从繁琐的手编程序的沉重劳动中解放出来，与此同时，学者们提出了这

种高级程序语言的形式描述，即巴库斯—瑙尔范式 (Bacus-Naur normal form，简称BNF)。后来发现，语言学家乔姆斯基 (N. Chomsky)的上下文无关文法(Context-free grammar，简称CFG)恰好与BNF等价，它们的数学形式在实质上是完全一致的，于是，BNF与CFG在数学上获得了高度的统一。因而乔姆斯基的工作引起了计算机科学界和数学界的广泛注意。由于这种数学上的高度统一，乔姆斯基的形式语言理论成为了计算机科学的基石之一，这一理论的提出，推动了计算机科学的发展。乔姆斯基在《自然语言形式分析导论》一文中，从数学的角度给语言提出了新的定义，指出:"这个定义既适用于自然语言，又适用于逻辑和计算机程序设计理论中的人造语言"。① 乔姆斯基在《文法的形式特性》一文中，专门用了一节的篇幅来论述程序设计语言，他讨论了有关程序设计语言的编译程序问题，这些问题，是作为"组成成分结构的语法理论的形式研究"，从数学的角度提出来的。他在《上下文无关语言的代数理论》一文中提出:"我们这里要考虑的是各种生成句子的装置，它们又以各种各样的方式，同自然语言的语法和各种人造语言的语法二者都有着密切的联系。我们将把语言直接地看成在符号的某一有限集合V中的符号串的集合，而V就叫做该语言的词汇……，我们把语法看成是对程序设计语言的详细说明，而把符号串看成是程序"。在这里，乔姆斯基把自然语言与程序设计语言放在同一平面上，从数学的角度，用统一的观点来加以考察，对"语言"、"词汇"等语言学中的基本概念，获得了高度抽象化的认识。他在《形式语法导论》一书的导言中指出:"生成语法的研究之能实现，乃是数学发展的结果，……普遍语法的数理研究，很可能成为语言理论的中心领域。现在要确定这些希望能否实现还为时过早。但是，根据我们今天已经懂得的和正在

① N. Chomsky, G. Miller, "*Introduction to the formal analysis of natural language*", 见《*Handbook on Mathematical Psychology*》(ed. by p. Lule et al), p283.

逐渐懂得的东西，这些希望未必是不合理的。"他乐观地预言："普遍语法的某种数学理论与其说是今日的现实，毋宁说是未来的希望。人们至多只能说，目前的研究似乎正在导致这样一种理论。在我看来，这是今天最令人鼓舞的研究领域之一，如果它能获得成功，那么，将来它可能把语言研究置于一种全新的基点上。"①

还有一种高级程序语言叫 ALGOL 60，这是一种用于科学计算的程序语言，ALGOL 60公布不久，人们在使用中发现了它存在二义性（即"歧义"），于是，计算机科学家们纷纷寻找机械的办法以便判断一种程序语言是否具有二义性，为此绞尽脑汁。后来，乔姆斯基从理论上证明，一个任意的上下文无关文法CFG是否有二义性的问题是不可判定的，由于CFG与程序设计语言的BNF等价，而ALGOL60的形式描述正是 BNF，因此，这种程序设计语言是否有二义性的问题也是不可判定的。乔姆斯基从 CFG与BNF 在数学上的一致性，有力地回答了计算机科学中的这一重大理论问题，充分地显示了数学对于语言学理论和计算机科学理论的作用。这样，也就吸引了许多有才能的数学家和计算机专家来关心语言学中的数学问题。

第四，机器翻译研究的深入以及立足于模式匹配的自然语言理解系统的研制，进一步推动了自动句法分析的研究，而这些研究都带有浓厚的数学色彩。

在语言学领域中，乔姆斯基提出了转换生成语法，韩礼德(M. A. K. Halliday)提出了系统语法，兰姆（S. M. Lamb）提出了层级语法，派克提出了法位学理论，盖兹达(G. Gazdar）提出了广义短语结构语法。这些语法理论都是相当形式化的，有着数学一般的严谨风格。

在计算机科学领域中，许多计算机专家和人工智能学者，也用数学方法来研究句法。伍兹(W. Woods）提出了扩充转移网络，

---

① M.Gross, A.Lentin, *Introduction to Formal Grammars*, 乔姆斯基的序言，Berlin，Springer-Verlag, 1970.

卡普兰（R. Kaplan）提出了通用句法生成程序，埃丁格尔(A. G. Oettinger)提出了预示分析法，凯依(M. Kay)提出了功能合一语法。这些理论和方法，都十分便于直接用于进行算法设计，便于在计算机上实现。

在这种情况下，出现了一大批兼通语言学、数学和计算机科学的人才，如语言学家布列斯南(J. Bresnan）和卡普兰提出的词汇功能语法，处处都使用了数学论证的方法。这种语法理论本身就是语言学和数学相互渗透而形成的绝妙产物。

传统句法学是用来教人学习句法分析的，而上述的各种带有数学风格和算法色彩的句法学则是用来教计算机进行自动句法分析的，当然也可以用它们来教人，这样的研究成果，进一步丰富了传统句法学的内容。

第五、语音的自动合成与分析是语言信息处理的一个重要方面。近三十年来，已研制出一批试验性的语音合成器，它们能把语音频谱转化为语音，这是十分困难的工作，因为语音频谱提供出来的信息实在是太多了，正如著名语音学家方特（G. Fant）所说的，人们很容易淹没在不了解其意义的各种声学特征的细节的汪洋大海之中。不过，从五十年代初以来，在语音合成器的研制方面仍然取得了有意义的成果。远在1939年，多德莱（H. Dudley）就在纽约的国际博览会上展出了“说话机”(talking machine)。1964年出现了肯佩棱机（Van Kempelen machine)，它能够自动合成大量的拉丁语、法语和意大利语词汇。这种语音合成器的研究，可以进一步揭示人类言语产生的机制，并可作为研究言语的产生和感知的工具。因此，目前国外在贝尔实验室、麻省理工学院、剑桥空军研究实验室、斯德哥尔摩皇家技术学院都进行过语音合成器的研究。我国在语音合成的研究方面已取得很大成绩，中国社会科学院语言研究所的汉语普通话语音合成，其自然度和逼真度都达到了“以假乱真”的地步。

语音分析的实质是把属于声学领域的连续的物理言语信号变

换为属于抽象语言学领域的离散的描述。奥登（K. W. Otten）曾指出，语音分析要注意四个主要问题：(1) 选择恰当的语言单元，(2) 把连续的信号转换为离散的信号，(3) 研究言语声学特征的可变性，(4) 研究言语的多余度。语音分析的结果就是语音的自动识别。国外已经研制成DRAGON HARPY等试验性的英语语音识别系统。我国在语音识别方面，主要围绕着特定说话者大词表语音识别系统和非特定说话者小词表语音识别系统开展工作。1986年哈尔滨工业大学研制出3000个孤立单词的语音识别系统。1988年清华大学利用矢量量化和隐式马尔可夫模型，研制成功能够识别30个城市名称的非特定说话者语音识别系统，1989年又研制出能识别 200 多个单词的实时非特定说话者语音识别系统。中国科学院声学研究所研制的2000个孤立单词的实时语音识别系统，在1988年西欧高技术展览会(TEC-88) 上获国际大奖，在此基础上已制成语音打字机。

语音的自动合成与分析牵涉到语音的语声统计特性、语言信号短期平均处理、频谱的分析与合成、短期傅里叶变换、语言的线性预测分析等数学问题，这是语言学与数学彼此协作、相得益彰的又一个天地。

第六，由于文字识别技术的进展，文字学研究开始同图象识别的方法结合起来。因为文字也是一种图象，图象识别中采用的许多方法，如图象识别的句法分析方法，也可用到文字识别中去，这方面的工作在美国、日本都取得了很大成就。图象识别的一般理论和方法也涉及许多数学问题，如何运用这些理论和方法来研究书面文字的结构，将是一个极有意义的新课题，也许会给古老的文字学研究开辟出一片新的天地。我国的汉字识别研究独具特色，采用选取汉字特征点和数学形态学的方法来提取汉字结构特征，目前，在印刷体汉字识别方面，已研制出一批实用系统，部分系统已经商品化，这些系统一般都具有版面分析、文本识别、识别结果后处理、自动纠错、自动编辑、自动输出等功能。在联

机手写体汉字识别方面，识别率正逐渐提高，已达到部分商品化的水平。在文字识别这个领域，数学也是大有用武之地的。

在电子计算机上输入输出英文、俄文等拼音文字（主要是拉丁字母和斯拉夫字母）的问题早已解决，但是象汉字这样的由数万个字符构成的大字符集，其输入输出计算机的问题就不是很容易的事了。为了解决汉字的输入输出问题，推动了汉字编码的研究，而为了使汉字编码研究建立在科学的基础上，就必须求助于数学，来研究汉字的频度，分析汉字的部件，测试汉字的信息量和多余度，设计高效率的汉字输入键盘。汉字的定量研究已引起了许多学者的关注。这是数学在语言学研究中发挥作用的另一个场所。

目前，在拉丁字母和斯拉夫字母以外的一些拼音文字，如泰文、朝鲜文、阿拉伯文、蒙文、藏文等在计算机上的输入输出问题，已有了可喜的研究成果，这些成果的取得在很多方面得益于文字统计特性的研究，在数学和计算机科学的推动下，传统的文字学研究园地里，吹起了一股现代化的东风。

第七，七十年代以来，建立了许多立足于语义的自然语言理解系统，使长期不受重视的语义学得到了发展，数学方法也在语义学中得到了应用。

近数十年来，不少语言学家认为，语义学不是语言学的一个分支，他们只关心语言形式的研究，而把意义的研究推给哲学或其它学科来进行。但是，随着机器翻译和自然语言理解研究工作的进展，再加上语言学理论论战的需要，促使语言学家去研究语义学。学者们逐渐认识到，甚至句法的研究也是不可避免地与语义学纠缠在一起的，因此，他们又重新对语义学发生了兴趣，并且这种兴趣很快地与日俱增。

哲学家们曾经提出过意义公设系统，它包括规则系统、蕴涵符号(——►)、逻辑连词(and、or、not)等，这样，便可以把词的意义分解为若干个基本意义组成的意义公设系统。例如：

boy⟶male

girl⟶female

man⟶male and adult

Woman⟶female and adult

boy or girl⟶not adult

female⟶not male

man or Woman or girl or boy⟶human

在意义公设系统中，词的意义可由一组语义公设来确定。哲学家们的这些研究，使意义获得可计算的性质，为用数学方法研究语义学打下了基础。在这种情况下，一些语言学家，如美国语言学家弗托(J. A. Fodor)和麦考利(J. D. McCawley)等，又把语言与逻辑的相互关系这样的问题重新提了出来。乔姆斯基关于深层结构和表层结构的理论，把语义问题提到了相当的高度，卡茨(J. Katz)和弗托等提出了解释语义学，采用成分分析法，利用语义成分、标记及关系来定义词符成分，并加上一些控制和选择限制来演绎地解释句子的语义。菲尔摩(C. J. Fillmore) 提出了格语法，从句子的深层句法表示来推导句子的表层结构。麦考利等提出了生成语义学，他们一开始就用语义结构来刻画句子，然后通过一系列的转换由这种语义结构产生出表层结构，而用不着对深层结构作任何说明。威尔克斯(Y. A. Wilks) 提出了优选语义学，并把这种理论用于英法机器翻译系统。在这些新的语义学理论中，都采用了数理逻辑的演算方法，充分地说明了数学对于语义学的深刻影响。

值得注意的是，有的数学家和计算机科学家也进行了语义学研究，他们也提出了一些有价值的语义学理论。如美国数理逻辑学家蒙德鸠 (R·Montague) 提出了蒙德鸠文法 (Montague grammar)，美国计算机科学家杉克 (R. C. Schank)提出了概念依存理论，美国人工智能专家西蒙(R. F. Simmons)提出了语义网络理论。这些理论出自自然科学家之手，简洁而严谨，直观而易

于操作，很快就在计算机上得到实现。在数学工具的推动下，语义学研究出现了空前活跃的局面。

总之，电子计算机的出现和广泛使用，就象催化剂一样促进了数学和语言学的结合。数学渗透到了形态学、句法学、词汇学、语音学、文字学、语义学等语言学的各个分支部门，促进了语言学的数学化。而语言学的数学化则是语言学现代化的一个重要内容，这些内容被概括在“数理语言学”(mathematical linguistics)这个新兴学科中，并得到了迅速的发展。

1955年,美国哈佛大学首先创办了数理语言学讨论班,1957年正式开设了数理语言学课程。接着，麻省理工学院、密歇根大学、宾夕法尼亚大学、印第安纳大学、加利福尼亚大学都相继开设了数理语言学课程。同年，日本成立了计量语言学会，创办了数理语言学杂志《计量国语学》，联邦德国的波恩大学也开设了数理语言学课程，苏联在莫斯科大学、列宁格勒大学及莫斯科国立第一外国语师范学院也进行了数理语言学的研究工作。1958年，莫斯科大学、高尔基大学、萨拉托夫大学、托姆斯克大学，分别给数学系及语文系的学生开设了数理语言学的选修课，并在列宁格勒大学设置了数理语言学专业。

此外，罗马尼亚、匈牙利、捷克斯洛伐克、英国、法国、挪威，德意志民主共和国、波兰、瑞典等国，都先后开展了数理语言学的研究工作，有的国家还创办了专门的刊物，成立了专门的研究机构。

我国从五十年代起便开展了数理语言学的研究工作。1982年，北京大学中文系给汉语专业的学生开设了《语言学中的数学问题》的选修课，首先在我国开设了数理语言学方面的课程①。1985年，上海知识出版社出版了我国的第一本数理语言学专著——《数理语言学》。数学的定量的研究方法已逐渐受到我国传统的语言学家

① 丁石孙、张祖贵，《数学与教育》，《数学·我们·数学》丛书，湖南教育出版社，1989年，第88页。

们的注意，并开始用到他们的研究工作中去，在用数学方法研究汉语的句子结构、汉字频率统计、汉语单词频率统计、频率词典的编制、方言定量分析、汉字熵值测定等方面，都取得一定的成绩。

数理语言学的研究常常要用电子计算机作为工具，因此，它与计算语言学的研究是联系在一起的。当前，数理语言学与计算语言学有合流的倾向。这清楚地说明，语言学、数学、计算机科学有着不解之缘。

十九世纪初叶，德国语言学家施来赫尔(A. Schleicher)把生物学中的分类方法用于语言发展过程的研究，提出了印欧系语言发展的谱系树，从而大大地推进了历史比较语言学的发展。二十世纪初叶，美国语言学家雅可布逊(R. Jakobson) 把物理学中关于物质由基本粒子构成的理论用于音位研究，提出了音位的区别特征学说，把音位学的研究发展到一个新的阶段。在当今的信息革命时代，把数学思想和数学方法用于语言研究，必将使语言学适应新技术革命的需要，从而促进语言学的发展，数理语言学有着广阔的发展前景。

数理语言学的产生和发展，强烈地冲击着从索绪尔以来的语言学理论，使我们更深入地了解到语言符号的许多重要特性，这些语言符号新特性的发现，必然从新的侧面进一步丰富我们对于语言符号本质的认识，加深我们对数学与语言的关系的理解。本书关于数学与语言关系的探讨，正是建立在我们对语言符号的新特性认识的基础之上的。

为了使本书的探讨有一个可靠的立足点，我们有必要谈一下我们对语言符号本质特性的认识。

瑞士著名语言学家索绪尔在他的《普通语言学教程》一书中，曾提出语言符号具有如下两个重要的特性：①

---

① 德·索绪尔，《普通语言学教程》，中译本，第103页，106页。

一、符号的任意性：语言符号的能指和所指的联系是任意的。索绪尔认为，符号任意性的原则“支配着整个语言学，它的后果是不胜枚举的，人们经过许多周折才发现它们，同时也发现了这个原则是头等重要的”。

二、能指的线条性，索绪尔指出，语言的能指属于听觉性质，只在时间上展开，而且具有借自时间的特征：（1）它体现为一个长度，（2）这长度只能在一个向度上测定，它是一条直线。索绪尔认为：“这是一个似乎为常人所忽略的基本原则，它的后果是数之不尽的，它的重要性与符号的任意性规律不相上下，语言的整个机构都取决于它”。

索绪尔提出的语言符号的这两个特性，当然是十分重要的。然而，索绪尔以后现代语言学的发展，特别是电子计算机出现以后现代语言学的发展，严竣地考验着索绪尔的理论。在我们看来，索绪尔提出的语言符号的任意性这一特性是无可非议的，但是，他提出的语言符号的第二个特性——能指的线条性就未必是正确的了，因为新的研究结果表明，语言的能指并不只是线条性的东西。英国著名语言学家弗斯（J. R. Firth）提出“跨音段论”(prosodic)，他认为，在一种语言里，区别性语音特征不能都归纳在一个音段位置上，例如，语调就不是处于一个音段位置上，而是处于前后相续的线条性的音段之外，笼罩着或管领着整个句子的东西。如果我们把语调这样的跨音段成份算进去，语言的能指就不宜于看作是线条性的东西，而应该看作是立体性的东西了。

索绪尔是一位出色的天才的语言理论家，他是名副其实的现代语言学的奠基人，他的语言学说，是语言学史上哥白尼式的革命，对于现代语言学的发展有着深远的影响。现代语言学的每一个部门，每一种流派，都直接或间接地受到了索绪尔语言学说的影响。他所说的语言符号的上述两个特性，是在当时的语言学和自然科学发展的水平下提出的。在索绪尔的时代，还没有电子计算机，数理语言学和计算语言学等新兴学科还没有形成，语言学主要是

与语言教学、文学、历史、考古学等学科有联系，在这种情况下，索绪尔当然不可能提出那些只有在电子计算机时代才能揭示的语言符号的新特点。随着电子计算机的出现和发展，语言学的理论也应该发展，我们决不能墨守陈规，满足于旧有的结论，而应该站在前辈学者的双肩上，高瞻远瞩，吸取电子计算机自然语言信息处理的新成果，结合现代数学的理论与实践，从新的角度，用新的眼光，以新的方法来研究语言这一极为复杂的符号体系。正是基于这样的认识，我们觉得，语言符号除了索绪尔所指出的那两个不尽完善的特点之外，还有着如下几个十分引人注目的特点：

1.语言符号的随机性：语言符号的出现和分布规律不是完全确定的，具有随机性，这一特性使得语言与统计数学发生了联系。

2.语言符号的冗余性：语言符号之间彼此制约，使得我们可以根据前后符号的关系来判断有关语言符号的性能，这样，语言符号就显示出冗余性，这一特性使得语言与信息论发生了联系。

3.语言符号的离散性：语言符号是由一些离散的单元构成的，具有离散性，这一特点使得语言与集合论发生了联系。

4.语言符号的递归性：语言符号可以反复地使用有限的规则构成无限的句子，具有递归性，这一特点使得语言与公理化方法发生了联系。

5.语言符号的层次性：语言的句子并不是由各个单词依前后的线性顺序排列而成的简单的线性序列，而是一个有层次的立体性结构，具有层次性。每一个句子的线性序列的表层之下，都隐藏着一个层次分明的树形图。这一特点使得语言与图论发生了联系。

6.语言符号的非单元性：语言符号并不是一个无结构的单元性符号，而是一个有结构的、由多个复杂特征构成的非单元性符号，具有非单元性，这一特点使得语言与数理逻辑的许多演算方法发生了联系。

7.语言符号的模糊性：语言符号中普遍存在着模糊现象，具有模糊性，这一特点使得语言与模糊数学发生了联系。

本书打算围绕语言的这些特性，来探讨数学与语言的关系。为了便于不同文化知识结构的广大读者理解和阅读，写作时尽量做到深入浅出，通俗易懂，以便引起对数学和语言学感兴趣的读者进一步来与我们探讨数学与语言的关系这一个问题，产生抛砖引玉的效果。

近年来，国内外在用数学方法研究语言方面取得了不少成果，本书力图反映出这些新成果，所引材料均在脚注中说明，作者谨对他们表示感谢。

由于本书牵涉到的知识面较广，作者所知极为有限，难免有不妥或错误之处，敬希读者指正。

第一章

语言符号的随机性与统计数学

第1节 语言符号的随机性

瑞士著名语言学家索绪尔在其名著《普通语言学教程》中把语言现象分为言语活动（langage）、言语（parole）和语言（langue）三样东西，它们之间是彼此联系而又相互区别的。

他指出，"言语活动是多方面的、性质复杂的，同时跨着物理、生理和心理几个领域，它还属于个人的领域和社会的领域。我们没法把它归入任何一个人文事实的范畴，因为不知道怎样去理出它的统一体。"① "因此，言语活动的研究就包含着两部分：一部分是主要的，它以实质上是社会的、不依赖于个人的语言为研究对象，这种研究纯粹是心理的；另一部分是次要的，它以言语活动的个人部分，即言语，其中包括发音，为研究对象，它是心理·物理的。"②

"把语言和言语分开，我们一下子就把(1) 什么是社会的，什

① 德·索绪尔，《普通语言学教程》，中译本，第30页。

② 同①，第41页。

么是个人的；（2）什么是主要的，什么是从属的和多少是偶然的分开来了。"①

他指出，"语言是一种表达观念的符号系统，因此，可以比之于文字、聋哑人的字母、象征仪式、礼节形式、军用信号等等，等等。它只是这些系统中最重要的。"②而言语则"是人们说话的总和"③，它包括言语行为的过程（也就是交际过程）和言语行为的结果（也就是口头的或书面的言语作品）。

索绪尔把语言比作乐章，把言语比作演奏，把语言和言语的关系比喻为乐章和演奏的关系。他说，"在这一方面，我们可以把语言比之于交响乐，它的现实性是跟演奏方法无关的；演奏交响乐的乐师可能犯的错误绝不会损害这种现实性。"④这是一个非常贴切的比喻。

在索绪尔关于语言和言语的区分的理论的影响下，美国语言学家乔姆斯基提出，必须把说具体语言的人对这种语言的内在知识和他具体使用语言的行为区别开来，并把前者叫做语言能力(competence)，把后者叫做语言运用(performance)。我们认为，乔姆斯基的语言能力，大体上相当于索绪尔的语言，乔姆斯基的语言运用，大体上相当于索绪尔的言语。

在言语(或语言运用)中，当我们用语言来进行交际的活动的时候，有的语言成分使用得多一些，有的语言成分使用得少一些，各个语言成分的使用有一定的随机性。当一个个的语言成分在话语和文句中出现时，有时它们的出现是完全确定的，有时则是不确定的。如果我们根据索绪尔和乔姆斯基的上述观点，把语言和言语区别开来，那么，我们就可以说，由于在言语中语言成分的使用和出现具有随机性，所以，我们可以采用统计数学的方法，

① 德·索绪尔，《普通语言学教程》，中译本，第35页。
② 同①，第37—38页。
③ 同①，第42页。
④ 同①，第40页。

从交际活动的角度和语言使用的角度来研究言语。

在有些场合，语言成分的出现还是确定的。例如，在俄语中，当浊辅音处于词尾停顿之前，一定要发生清化的现象。这种浊音清化的现象是完全地确定的。

способ(方法)——[sposəp]　　b→p

завод(工厂)——[zavot]　　d→t

нож(刀子)——[noʃ]　　ʒ→ʃ

浊辅音b、d、ʒ变成了相应的清辅音p、t、ʃ。

如果我们把“浊辅音处于词尾停顿之前”看成一个条件组，而把“浊音清化”看成这个条件组的结果或简称之为事件，那么，我们可以说，当实现了条件组“浊辅音处于词尾停顿之前”时，一定会发生“浊音清化”这一事件。这种事件，叫必然事件。抽象地说，如果实现了某一确定的条件组S，就一定会发生与之相应的完全确定的事件A，则A就叫做必然事件。在语言交际活动中，这样的必然事件是极为少见的。

在语言交际活动中出现的大量事件并没有这种完全的确定性，通常都有例外。例如，对于“英语中名词之前出现定冠词”这一事件A，我们就不能穷尽地找出单义地确定定冠词出现的条件组S。因此，当我们实现某个条件组S时，“出现定冠词”这一事件可能发生，也可能不发生。

如果当S＝“名词是一个季节名词”时，我们可以看到在“Spring has come”（春天来了)中，不出现定冠词，而在“in the spring”(当然也可以用spring，意思是“在春天”)中，出现定冠词，也就是说，“出现定冠词”这一事件可能发生，也可能不发生。

如果当S＝“名词是专有名词”时，我们可以看到，在Shanghai(上海)，Britain(不列颠)、John Brown(约翰·布朗)中，不出现定冠词，而在the Yellow River(黄河)，the Baltic Sea(波罗的海)、the pacific Ocean(太平洋)、the Himalayas(喜马拉雅山)中，出现定冠词。

当实现了条件组S时，某一事件A可能发生，也可能不发生，这种事件叫做对于该条件组的随机事件。

由于在交际活动中，当实现了某一条件组时，语言成分的出现在绝大多数场合是不确定的，是有例外的。因此，我们可以说，在语言交际活动中，语言成分的出现是一个随机事件。从本质上说来，语言符号具有随机性，这样，在交际活动出现这样的随机事件便是很自然的了。

正因为语言符号具有随机性，因而很难用确定性的规则来描述它。几乎每一条语法规则都有例外，这种例外现象使得研究语法的语法学家们伤透脑筋，有的语法学家甚至为此而误入迷津，以偏概全，得出了错误的结论。为了避免以偏概全的错误，我国前辈语言学家曾提出“例不过十不立，反例不过十不破”的原则来制定语法规则，这个原则常常作为判断语言学家治学态度是否严谨的准绳。其实，对于言语活动这样随机现象来说，找出十个例子来立某条语法规则并不难，而找出十个反例来破某条语法规则也很容易，以十个例子或十个反例来作为某条语法规则破或立的标准，看来未必恰当。最好的办法还是采用统计数学的方法来对交际活动中所出现的各种语言现象进行描述。如果我们能够从理论的高度，把随机性看成是语言符号本身的一种自然特性，并采用恰当的数学工具来描述这种随机性，那么，我们对于语法规则中的大量的例外情况也就不会再感到迷惑不解和束手无策了，因为这些例外情况正是由于语言符号本身的随机性这一特点而形成的。

事实上，在语言成分的出现这一个随机事件中，随机事件A与条件组S之间虽然没有完全确定的联系，但是，它们之间却有着统计上的联系。尽管当条件组S实现一次时，事件A可能发生，也可能不发生，但是，如果条件组S实现多次，事件A的发生就有着某种规律性，这种规律性表现为事件A发生的频率。①

① 在我国，许多从事言语统计研究的学者又把频率叫做频度。本书中一般叫做频率，只在引用有关资料时，如原资料中叫频度，才用“频度”这个术语。

所谓频率，就是事件A实际出现的次数与条件组S实现的次数之比，可用下面的公式表示：

$$f=\frac{n}{N}$$

其中，f表示频率，n是事件A的实际出现次数，N是条件组S的实现的总次数。

例如，在英语中，当条件组“是季节名词”实现500次时，有400次季节名词前不带定冠词，那么，“季节名词前不带定冠词”这一随机事件的频率为：

$$f=\frac{n}{N}=\frac{400}{500}=0.8=80\%$$

当条件组S的实现次数较少时，随机事件A发生的频率是不稳定的，有时发生的频率高些，有时发生的频率低些，但是，当多次实现条件组S时，随着实现次数的增加，随机事件发生的频率越来越稳定于一个确定的值，这种当条件组S多次实现时，随机事件发生的频率渐趋稳定的规律性，与前面所说的完全确定的规律性不同，我们把它叫做统计规律性。

例如，在翻译中，当我们采用权威性的英语语法著作中关于定冠词的配置的规则来把汉语译为英语时，我们在定冠词的使用上有时会发生错误，也就是说，虽然条件组S（即英语语法书中关于定冠词的配置规则）实现了，而事件A（即相应定冠词的选择）却并不发生。但是，当我们把这样的规则用来翻译大量的英语资料时，如果我们采用的英语语法著作编得确实好，那么，我们会发现，在大多数情况下是能正确地选择定冠词的。比如说，在100个场合有80个场合选择定冠词是正确的，那么，这种选择定冠词的规则就是统计规则。

令人可喜的是，近年来在我国的语法研究中，不少语法学家开始认识到语言符号的这种随机性，自觉地采用统计数学的方法来描述汉语语法现象。

汉语中有一类可能补语有肯定和否定两种形式，这种可能补语是在动词和结果补语或趋向补语之间插入"得"或"不得"构成的。插入"得"构成肯定形式，插入"不得"构成否定形式。例如：

吃饱——吃得饱——吃不饱

出来——出得来——出不来

虽然这种可能补语有肯定和否定两种形式，但在实际语言中主要用否定形式，较少用肯定形式。我国学者根据《曹禺剧作选》、《骆驼祥子》、《老舍剧作选》、《李有才板话》、《李家庄的变迁》和《李自成》(二卷下) 1 145 000字的材料仔细地作了统计，发现这种可能补语否定形式与肯定形式之比为1211:42。否定形式的出现次数为肯定形式出现次数的29倍。①

当我们表达主、客观条件不容许实现某种结果或趋向时，一般很少用"不能+动词+结果补语（或趋向补语)"的语法格式，而采用可能补语的否定形式。例如：

① 吸烟的害处说不完。

  *吸烟的害处不能说完。

② 银花想不出办法来。

  *银花不能想出办法来。

其中，标有"*"号的句子是不能说的。

这样的语言事实，可以说明为什么这一类可能补语的否定形式用得远比肯定形式高29倍。

当用"能"和"可以"表示"主客观条件允许"的意义时，在肯定形式里可以互相替换。例如：

③ 小明一口气能跑五十米。

  小明一口气可以跑五十米。

但在否定形式里主要用"不能"，极少用"不可以"。例如：

④ 小明一口气不能跑三千米。

①刘月华，《可能补语用法的研究》，(《中国语文》)，1980年，第4期。

*小明一口气不可以跑三千米。

统计表明，在这种语言格式中，"不能"与"不可以"的出现次数之比为148:8。这8例"不可以"，均出自文言色彩较浓的作品中，根据这样的统计，可以得出这样的规律：现代汉语中一般只说"不能"，它的出现频率为0.95。

我国学者还用统计方法研究了状语中用"地"的情况。根据曹禺的《雷雨》、老舍的《骆驼祥子》、谌容的《人到中年》、宗璞的《三生石》以及《中国共产党中央委员会关于建国以来党的若干历史问题的决议》约437 000字中状语用"地"的情况，发现描写动作者的状语大多数用"地"，用与不用之比为1158:66，描写动作的"地"往往可用可不用，用与不用之比为675:2273。

例如：

⑤ 他不动声色地一件件处理着。

在这个句子中，"不动声色"是描写动作者的，故用"地"，"一件件"是描写动作的，故不用"地"①。

根据这样的理由，我们把描写性状语分为描写动作者的与描写动作的两种。

我国学者在研究北京话的拟声词时，发现由四个不同音节构成的A B C D式的拟声词与双音节拟声词有对应关系。例如：

A B C D式拟声词	双音节拟声词
哔滴叭哒	哔叭（AC）
	叭哒（CD）
	滴哒（BD）
丁零当啷	丁当（AC）
	当啷（CD）
	丁零（AB）
乒零梆啷	乒梆（AC）

① 刘月华，《状语的分类和多项状语的顺序》，（《语法研究和探索》）（1），第38页。

榔啷（CD）

唧里咯登　　　　　　　　　　咯登（CD）

因此，他们认为，ABCD式拟声词是由双音节拟声词变来的。

但是，究竟ABCD式是从哪一种双音节形式变来的？还是不同的词有不同的变化？为此，对101个ABCD式拟声词进行了统计，发现这101个ABCD式拟声词都有相应的CD式双音节形式，而AC式、BD式、AB式等双音节形式则少得多。统计结果如下：

表1.1.1

形　式	ABCD式	AC式	CD式	BD式	AB式
数　目	101	39	101	18	3

这样的统计数字证明，ABCD式和CD式之间的关系最为密切。大部分ABCD式没有相应的AC和BD形式，至于AB形式则更少。因此可以认为ABCD式是CD式的一种变化形式，而不是AC式或BD式的变化形式，更不是AB式的变化形式。①

当某种语法现象有不止一种意义或用法时，如果不通过大量的语言材料进行统计调查，凭感觉判断而仓促作出结论，就可能忽略掉一些能反映重要规律的现象，或者把有规律性的随机的语言现象错误地当作例外加以处理。运用统计方法可以避免这方面的疏漏，或者发现过去在这方面的疏漏，得出比较客观的结论。

有人曾研究过俄国诗人普希金（Пушкин）、屠格涅夫（Тургенев）和蒲宁（Бунин）的诗歌中动词Быть的出现的情况。在这种研究中，条件组S＝"普希金、屠格涅夫和蒲宁的诗歌"，事件A＝"Быть出现"。

当条件组S实现次数很少时，文句容量为10个词，Быть的出现次数为0，Быть的出现频率当然也是0；当条件组S的实现次数稍增，文句容量为100个词时，Быть出现3次，Быть的出现频率为0.030；当条件组S的实现次数继续增加，文句容量逐渐加大，

① 孟琮，《北京话的拟声词》，(《语法研究和探索》(1))，第120页。

Быть的出现频率越来越稳定，最后稳定于0.010左右。Быть 出现频率的变化情况，如下表所示：

表1.1.2

文句容量	10	100	1000	2000	3000	4000	5000
出现次数	0	3	15	17	31	33	47
出现频率	0.000	0.030	0.015	0.008	0.010	0.008	0.009
文句容量	6000	7000	8000	9000	10000	15000	40000
出现次数	57	71	74	88	95	153	4186
出现频率	0.010	0.010	0.009	0.010	0.010	0.010	0.011

我们曾研究过汉字中常用汉字“的”字随着文句容量的增大其出现频率的变化情况。在这种研究中，条件组S=“汉字文句”，事件A=“‘的’字出现”。

当条件组S实现次数较少时，文句容量为15215个汉字，“的”字出现776次，出现频率为0.051；当文句容量为80125个汉字时，“的”字出现3365次，出现频率为0.042；当条件组S的实现次数继续增大，文句容量逐渐增大，“的”字的出现频率越来越稳定，最后逐渐稳定于0.042。这种情况，如下表所示：

表1.1.3

文句容量	15215	80125	813526	1125370	1429452	5239153
出现次数	776	3365	33355	46140	60037	220044
出现频率	0.051	0.042	0.041	0.041	0.042	0.042

由这些例子可以看出，只有多次实现条件组S，才有可能建立随机事件A的统计规则。设条件组S的实现次数为t，随着t的增大，随机事件A的出现频率f越趋稳定，当$t\to\infty$时，f就越近于一个定值，这个定值叫做随机事件A的概率，记为p，即

$$\lim_{t\to\infty} f = \lim_{t\to\infty}\frac{n}{N} = p$$

从公式可看出，由于n不大于N，因而随机事件的概率总是正数，并且总是处于0和1之间，即

$$0 \leqslant p \leqslant 1$$

如果$p=0$，则随机事件是不可能事件。

如果$p=1$，则随机事件就变成完全确定的事件，即必然事件了。可见，必然事件只不过是随机事件当$p=1$时的一种特殊情况。所以，在语言交际活动中，语言成分的出现不论是完全确定的也好，不完全确定的也好，都可以看成随机事件。如果是完全确定的事件，那就可以是随机事件当$p=1$时的一种特殊情况。正是在这个意义上，我们才把语言符号的随机性看成是语言符号的一个普遍性质。

第2节　字频和词频的统计

目前世界上的语言共有2500种至3000种。其中美洲语言多达1000多种，非洲语言也近1000种。语言使用人口超过100万的只有140种。其中，使用汉语的人最多，占世界人口的20%，其次是英语，约3亿人口；再次是俄语、西班牙语和印地语。使用上述这五种语言的人共占世界人口的45%，再加上使用日语、德语、阿拉伯语、葡萄牙语、法语和意大利语的人，占世界人口的60%。

在世界上这么多的语言中，有的语言没有书面形式，只有口头形式。书面形式的语言要用文字来记录。当今世界的文字，有的地区用汉字，有的地区用字母。

用汉字作正式文字的国家有中国、日本(汉字假名混合使用)、朝鲜（北方全用谚文，南文汉字谚文混合使用）和新加坡（同时

以英文、马来文和泰米尔文为官方文字)。

字母有多国通用的，有一国独用的。三种多国通用的字母占去了地球的一大半。①

分布最广的多国通用字母是拉丁字母，又叫罗马字母。欧洲的大部分、美洲和澳洲的全部、非洲的大部分和亚洲的小部分，都以拉丁字母作为正式文字。欧洲有一条字母分界线，沿着苏联和保加利亚的西面边境，穿过南斯拉夫的中部；分界线以西是历史上的罗马天主教国家，用拉丁字母；分界线以东是历史上的东正教国家，用斯拉夫字母。非洲也有一条字母分界线，这就是非洲北部的几个阿拉伯国家的南面边境，分界线以北用阿拉伯字母，以南的大半个非洲用拉丁字母，新独立的国家大都沿用原宗主国的拉丁字母文字（英文、法文、葡萄牙文等），也有当地语言写成拉丁字母的，例如，在南非有阿非利根斯文，在东非有斯瓦希里文等。亚洲用拉丁字母作为正式文字的国家有土耳其、越南、印度尼西亚、马来西亚、菲律宾和新加坡（以英文为主要文字）。

第二种多国通用字母是阿拉伯字母，它的分布仅次于拉丁字母。阿拉伯字母是北非和西亚十几个阿拉伯国家的正式文字。此外，有些伊斯兰教国家和地区也用阿拉伯字母作为正式文字，如伊朗、阿富汗、巴基斯坦、中国的新疆维吾尔自治区等。

第三种多国通用字母是斯拉夫字母。斯拉夫字母是苏联俄罗斯族和少数民族的文字（波罗的海东岸的三个加盟共和国爱沙尼亚、拉脱维亚、立陶宛除外）。保加利亚、半个南斯拉夫（另一半用拉丁字母）、蒙古人民共和国也以斯拉夫字母为正式文字。

除了这三种多国通用字母之外，还有印度字母系统。这种字母系统包括多种字母，同出一源而形体各异，不能彼此通用。印度字母系统用于印度各邦（有主要的14种文字）、印度四邻(斯里兰卡、孟加拉、尼泊尔、不丹、中国的西藏)、印度支那半岛（缅

① 周有光，《文字的国际分布和历史演变》,(《语文建设》)，1988年，第5期，第16页。

甸、泰国、老挝、柬埔寨)。

此外，一国或一地区独用的字母主要有：希腊字母(希腊)、希伯来字母（以色列)、假名字母(日本)、谚文字母（朝鲜)、阿姆哈拉字母（埃塞俄比亚）和蒙古字母（中国的内蒙自治区)。

语言使人类别于野兽，文字使文明别于野蛮。为了让人们能够读书识字，在人类的历史上，很早就开始了制定词表的工作。据说3000多年前，古代巴比伦的楔形文字里就保存有最早的词表。盎格鲁一撒克逊学者艾尔弗利克（Aelfric）写于9世纪的《拉丁语法》里，也包括了一份拉丁语-英语分类词表。18世纪法国莱贝神甫（abbé de l'Epée）制定了一份以1800个词为一个阶段、三个阶段共5400个词的词表，用于聋哑儿童教学。词表的制定是言语统计的一个古老部门，用统计方法编制词表，可以从语言现象的量的描述得出质的评价。1898年，德国学者凯定（F. W. Kaeding）编制了世界上第一部频率词典《德语频率词典》，这部频率词典的编纂是建立在大量的词汇统计工作的基础之上的，其目的在于教授速记学，创造新的速记方法。

二十世纪以来，随着国际交流的日益频繁，外语教学得到了蓬勃的发展。各国语言学家纷纷从事为语言教学服务的基础词汇表的研究和制定工作。20世纪前70年间，单是德语词表就至少有60种以上，而英语和法语词表的数目也不会少于此数，西班牙语词表和俄语词表也很多。仅这5种欧洲语言的词表，就有三、四百种之多。许多词表都是在词汇统计的基础上制定的。

1949年，法国学者米谢阿（R. Michéa)提出要建立词汇统计学，他认为这将是“一门年轻而富有前途的科学”。1954年，法国学者基罗（P. Guiraud）根据文章中词的频率分布提出了词汇丰富度的概念，他又于1960年出版了《统计语言学的问题和方法》一书，德国学者哈特曼（R. Hartmann）认为语言现象的统计研究可以叫做“语言学中的统计方法”，又可称为“词汇频率研究”。1956年，英国统计学家赫尔丹（G. Herdan）发表了《语言是选择

和机遇》*Language as Choice and Chance*)一书，系统地总结了语言现象统计研究的成果。1965年，德国学者凯尔（R. D. Keil）把词频统计和现代统计学结合起来，提出了"词汇计量学"(lexicometric)。

根据索绪尔语言和言语区分的理论，词汇计量学中所研究的语言现象的统计规律，都应该属于"言语"的范畴，因此，许多学者又把这样的研究称为"言语统计"。

近30年来，由于在言语统计中广泛地采用电子计算机，逐渐地改变了手工查频、手工统计的方法，提高了统计的效率和精度，把言语统计的研究提高到了一个新的水平。苏联拉脱维亚共和国科学院语言和文学研究所数理语言学实验室，运用电子计算机对现代拉脱维亚语的词汇、构词和形态进行了大规模的言语统计，分析了120万词的资料，编出了四卷本的《拉脱维亚语倒序频率词典》，对名词、形容词、动词的后缀作了统计描述，指出了每个后缀的总频率数值及其在各类文章中的分布情况。该实验室还用计算机对词类、词类的范畴及形式进行统计描述，计算了不同词类在各类文体文章中所占的百分比，这样大规模的言语统计，是传统的手工统计方法很难做到的，充分地显示了电子计算机对言语统计的巨大威力。

言语统计的一个重要目的，是为词表的制定提供统计方面的根据。因为既要制定词表，就必须选词。不选词，词表的制定就无法进行。选词标准有两类：一类是主观标准，一类是客观标准。①

主观标准也叫经验标准。词表的编制者根据个人的学识、经验和兴趣，来判断词表中应该收哪些词？不应该收哪些词？历史上的不少词表都是以主观经验为标准制定的。

30年代初期，英国学者奥格登（C. K. Ogden）和理查兹(I.

---

① 程曾厚，《"词汇计量学"的三项选词标准》，(《语文现代化》)，1983年，第1辑，第40页。

R. Richards）以主观经验为根据提出了“基础英语”(Basic English）词表。这份词表只有850个英语单词，他们宣称，只要用这850个单词，就可以给语言中的一切词下定义，从而用这850个数量有限的单词来表达人类思维活动的一切内容。

主观的选词标准根据的是专家的经验和断判力，如果专家水平很高，他们提出的选词标准诚然也有其可取之处。不过，人们更倾向于使用客观的选词标准。因为客观的选词标准是根据言语统计的方法得出来的，统计结果不以个人的意志为转移，这样制定出来的词表才会有较高的科学价值。

迄今发现的客观选词标准，除了频率之外，还有分布率、易联想性、扩散率、覆盖率、通俗性等。其中最主要的有 3 项，即频率、分布率和易联想性。下面分别加以说明。

（1）频率标准　大规模地根据频率来进行词汇工作是从德国学者凯定的《德语频率词典》开始的。凯定为了创制一种合理的速记法，对德语进行词汇频率统计。他准备了110份词汇材料，每份词汇材料包含10万左右的词汇，总词汇量达10 910 777个单词。这些材料主要取自报刊杂志（占总词汇量的40%）等14个不同的领域，从中抽取了出现次数在 4 次以上的不同词共79716个。全部统计工作用手工方式进行，动员了 5000 名速记人员和800名合作者参加。

《德语频率词典》的编制为德语基础词汇的研究打下了良好的基础。凯定搜集的词汇材料内容比较广泛，各部分内容的比例也比较得当，统计结果有一定的代表性。

但是，频率标准有局限性。因为频率指数的可靠是相对的。假定有10篇词汇量大致相等的语言材料，有A、B两个词，词A的出现频率是0.020，词 B 的出现频率是0.018，如果词A只在一篇语言材料中出现，而词B在10篇语言材料中都出现，显然词B比词A更为有用，尽管词B的频率比词A的频率稍低，但它在语言材料中的分布比词A 广泛得多。可见，频率有时会掩盖事物的真相。

法国学者基罗在谈到频率的相对性时指出："整理一批300 000个词的语言材料时，词表超过第500个词以上，得出的结论就不再可靠了；有1 000 000个词，才能定出一张大约1 000个词的词表来；而一张大约有2 000～3 000个词的词表，要以5 000 000个词的样本作为基础。"①

（2）分布率标准 时代的不同，地域的不同，材料长度的不同，材料篇数的不同，材料是口头语言还是书面语言，这些因素都会影响到语言材料的内容，从而使得我们不能只以频率标词作为选择词汇的唯一标准。

一个词在一定篇数的语言材料的样本中出现在多少篇数中，也是衡量该词重要与否的标准。这个标准，叫做分布率(range)标准。

加拿大学者贝克（E. Varder Beke）于1935年出版的《法语词汇手册》一书中，引进了分布率标准来进行词汇研究。他统计的语言材料共88篇，总词汇量在1 100 000以上。每篇材料原则上为10 000词，实际上平均为13 000词。88篇材料中，题材广泛，其中，小说、故事共34篇，剧本12篇，占总词汇量的56.4%。选词时以分布率为主要标准，其次才考虑频率标准。贝克认为，一个词如果有五位作家各用一次，也比另一个只被一位作家使用十次的词更重要。贝克的统计工作也是手工进行的，历时几近一年。所统计的88篇材料共收不同的词19 253个，词表中只收分布率指数为5以上的词，共6067个。分布率指数在5以下的词有13186个，占总词数的68.5%，尽管其中有些的频率很高，只因分布率指数在5以下，也被淘汰了。

贝克如此强调分布率的作用，未免有些失之偏颇，不过，他给那些只以频率作为唯一选词标准的学者敲起了警钟，使他们认识到，分布率标准也是不可忽视的。法国学者缪勒（C. Muller）

① P. Guiraud, *Problèmes et méthodes de la Statistique linguistique*, p155-156.

说得好："频率概念如果不立即与分布率概念相结合，那么，频率概念的价值是不高的。"①

（3）易联想性标准　有些表达具体事物的具体名词，在日常生活中十分有用，但它们表现出来的频率和分布率都很低。如果只根据频率标准和分布率标准来选词，那么，这些十分有用的具体名词就会被忽略了。

这些具体名词的频率和分布率虽不高，但它们随时都能使用，一有需要便立即在脑海中出现。因此，学者们提出了选词的第三个标准——易联想性标准（availability）。最早提出易联想性标准的是法国学者米谢阿，他指出："易联想的词频率并不特别高，但是它们随时都能使用，一有需要便立即在思想中出现。"②

易联想性按主题来进行调查，被调查者就某个主题写出他最先联想到的那些词，再从搜集到的全部易联想词中，按频率选出名列前茅的词。

法国学者古根内姆（G. Gougenheim）、米谢阿、里旺克(P. Rivenc)、索瓦若(A. Sauvageot) 采用了易联想性标准，于1954年完成了《基础法语》(Frangais fondamental) 的词表。

《基础法语》词表的研制，由法国政府提供经费，委托著名学者合作进行，并专门设立了"基础法语研究中心"(1959年改名为"法语传播研究中心"，现在是常设机构)。词表的调查工作从50年代初期开始，1954年7月出版《基础法语》，1956年出版《基础法语的制定》一书，阐明《基础法语》制定的原则、方法和过程。

基础法语调查在现场采录了163篇访问材料、访问时，要考虑被访问者在地区、年龄、性别、文化程度方面的差别。全部材料共包括312 135个词，不同的单词7 995个。

---

① C. Muller, *Quelques méthodes d'analyse du vocabulaire*, p164.

② R. Michéa, *Limitation et séelection du Vocabulaire dans lénseignement actif des langues vivantes*, «*Revue de langue vivantes*», No 22, p467.

《基础法语》词表根据使用频率、分布率和易联想性三项客观标准来选词。首先，从7995个不同的单词中选出出现20次以上的词1063个，编出频率词表，他们是把出现次数当作频率的，这种频率是绝对频率，而不是相对频率。然后，从1063个词的频率词表出发，经过如下两步复杂的筛选过程，才得到包含1475个词的正式词表。

第一步：从1063个词中留取出现次数29次以上、分布率指数5以上的词，以淘汰频率不十分高以及那些由于偶然原因而导致高频率的词。这样，1063个词减少到805个词，然后，再从中剔除104个被认为不宜于进入基础词表的词，如"俗词"等，这样，只剩下701个词，这些词是根据绝对频率和分布率指数两个标准选出的。

第二步：在701个词的基础上，进一步补充易联想词。他们把9—12岁的小学生分成若干组，就16个"主题"来调查易联想词，16个主题中的前5个主题是：身体部位，衣着，住房，家具，食品和饮料。调查者请小学生针对每个主题写下最先想起来的20个名词。把每个主题搜集到的词按频率标准选出其中的高频率词，这些词就是易联想词。把这些易联想词补充入原先选出的701个词中，共得1475个词。

在《基础法语》词表的1475个词中，实词有1222个，虚词有253个。实词中，名词有692个，占46.9%，其中大部分是以易联想词入选的具体名词；动词有339个，占22.9%。253个虚词占17.1%。与原来频率词表中的1063个词相比较，名词从395个增加到692个，所占的百分比从37.16%上升到46.9%。具体名词数量的增加，使得整个词表中各种词的比例更加合理。

按同样的方法，他们又选出《基础法语》的第二批词表，前后两批词表分别称为第一阶段词表和第二阶段词表。1958年古根内姆出版的《基础法语词典》中，包括了这两个阶段的全部词汇。

另一个有代表性的工作是美国普菲费尔（J. A. Pfeffer）主

持的"基础德语"(Grunddeutsch) 词表。

普菲费尔于60年代从德国到美国,在美国有关机构的资助下,成立了基础德语研究所,1964年,发表了《基础德语(口语)词表(基础阶段)》。

普菲费尔从德国、奥地利、瑞士等国的德语区挑选了70个城市和乡村,收集了6 00 000个词的口语材料,又从中学生中,根据有关主题选取了833 000 个词作为易联想词的材料。他也使用了频率、分布率和易联想性三项标准,整理出一份1084个词的词表,又根据经验加上作者认为非加不可的185个,编成包括1269个词的《基础德语词表》。

选词工作可分为三步。第一步以绝对频率和分布率指数为标准,选取出现次数大于40、分布率指数大于25的口语词737个。第二步选易联想词。由15—16岁的中学生就25个主题提供包括名词、动词和形容词在内的易联想词,再从中留取分布率较大、出现次数在100次以上的易联想词347个,前后两步共得单词1084个。第三步是根据作者的经验补选185个词,如表中有"太阳",就补选"月亮"和"星星"等。这185个词有3/4已直接或间接地在频率词表和易联想词中出现过,真正根据经验加进的词不足50个。

1970年,普菲费尔出版了《基础德语(口语)词表(中级阶段)》,并出版了《日常使用基础德语(口语)词典》。

两种基础词表收词都在1000个左右,这反映了法语和德语两种语言在词汇结构上的一致性。法国学者拉加纳(R. Lagane)认为:"超出最常用词的核心(至多1000个)之外,频率的概念就没有多大意义了。"①看来频率在词表的制定中起着重要的作用,它是选词的基本标准,但是,我们不可过分夸大它的这种作用。

我国是使用汉字的国家,因此,我国学者在词表方面的研究,是从汉字的频率统计开始的。

---

① R. Lagane, *De la notion de vocabulaire essentiel*, «*Grand Larousse de la langue française*», *Tome I*, L××××11～L××××111.

汉字是一个大字符集。目前世界上的表音文字，其字符数目都很有限，拉丁字母26个，斯拉夫字母33个，阿尔明尼亚字母38个，塔米尔字母36个，缅甸字母52个，泰文字母44个，老挝文字母27个，藏文字母35个，朝鲜文字母24个，日文假名48个。但是，汉字的字符成千上万，很难说出其准确数目。

我国的第一部字典，东汉时代许慎编著的《说文解字》，共收单字9352个，异体字1163个。晋朝吕忱编著的《字林》，收字12824个。南北朝顾野王编著的《玉篇》，收字16917个。宋朝陈彭年等编著的《广韵》，收字26194个，丁度等编著的《集韵》，收字53525个，王洙等编著的《类编》，收字53165个。明朝梅膺祚编著的《字汇》，收字 33179 个。清朝陈廷敬等编著的《康熙字典》，收字 47043 个(增补前为42174个)。1915年欧阳溥存等编著的《中华大字典》，收字 48000 多个，1959年日本诸桥辙次编著的《大汉和辞典》，收字49964个，1971年张其昀主编的《中文大辞典》，收字49888个。随着时代的推移，字典中所收的字数越来越多。最近开始分册出版的《汉语大字典》，所收的字数将超过56000字。可见，汉字确实是一个相当庞大的字符集。

我国早在20年代就开始进行汉字的频率统计，汉字的频率叫做字频。首先不是进行词频而是进行字频统计，这是我国言语统计研究的一个特点。著名教育学家陈鹤琴先生为了教学的目的编写了《语体文应用字汇》，于1925年完成，于1928年由商务印书馆出版。陈书前有《绪论》，叙述“中文应用字汇”曾有多种，其中包括克仑茨 (Pastor P. Kronz) 的研究和他自己编写的《常用四千字录》。陈鹤琴做过两次统计，第一次统计使用六种材料包含554 478个汉字的语料，得不同汉字4 261个；第二次使用34 818个汉字的语料，得出与4 261字相异的不同汉字 458个。第二次统计所得成果毁于火，在《语体文应用字汇》中印出的只是第一次的统计成果。

陈鹤琴用的语料分六类：

1.儿童用书：127 293字；

2.报刊（以通俗报刊为主）：153 344字；

3.妇女杂志：90 142字；

4.小学生课外作品：51 807字；

5.古今小说：71 267字；

6.杂类：60 625字。

书末附有"字数次数对照表"，即按汉字的绝对频率排列的字表。

我国著名教育学家陶行知先生为《语体文应用字汇》写了序言。序言中说："他们（指"近代教育家"）对于一门一门的功课，甚至一篇文章，一个算题，一项运动，都要依据目标去问他们的效用。他们的主张是要所学的，即是所用的。……到了后来他们连学生学的字也要审查起来了。学生现在所学的字，个个字都是有用的字吗？自从这个问题发生就有好几位学者开始研究应用字汇。我国方面也有几位先生研究这个问题，其中以陈鹤琴先生的研究为最有系统。他和他的助理九人先后费了二三年工夫，检查了几十万字的语体文，编成这册《语体文应用字汇》。这册报告未付印以前已经做了《平民千字课》用字的根据。将来小学课本用字当然也可以拿他来做一个很好的根据。虽然不能十分完备，但我想这本字汇对于成人及国民教育一定是有很大的贡献的。"①

1946年8月，四川省教育科学院根据陈鹤琴的《语体文应用字汇》和杜佐周、蒋成堃的《儿童与成人常用字汇之调查与比较》，按照两种字表相加后绝对频率的多少，选出最常用的字2 000个，编成《常用字选》。上述两种字表统计语料的总字数为775 832个。

新中国成立后，国内不少单位用手工做过《毛泽东选集》用字统计，据云南冶金第三矿统计，《毛泽东选集》1—4卷简体普及本用字总数660 273个，使用不同汉字3 002个。

---

① 陈鹤琴，《语体文应用字汇》，商务印书馆，1928年。

台湾省交通大学花了2 000 多个人日，根据200余万字的资料，也进行过汉字频率的统计工作。

1974年8月，原四机部、一机部，中国科学院、新华通讯社联名向国家计委申请研制"汉字信息处理系统工程"，同年9月，国家计委下文，批准这一工程，并提出，这一工程由四机部组织领导，成立领导小组和办公室。这就是有名的"748工程"。

研制汉字信息处理系统，首先要弄清汉字的属性和使用情况，进行汉字统计研究，以便为"748工程"提供数字依据。为此，"748工程"领导小组和国家出版局商定并拨出专款，开展汉字频率的统计研究，由北京新华印刷厂和北京市印刷技术研究所等19个单位参加，用两年的时间，把从各单位收集来的三亿多字的出版物，分成科学技术、文学艺术、政治理论和新闻通讯四类，并从中选出86本书、104本期刊、7 075篇论文，合计21 657 039个字，作为统计研究的样本，四类语料同时进行频率统计，最后汇总成一份综合资料，提供"748工程"使用。他们的统计是用手工进行的，从21 657 039个汉字样本中，统计出不同的汉字为6347个，并编成了《汉字频度表》。①

但是，手工查频费时费力，容易出错，统计样本范围越大，字数越多，出错率就越高。例如，《毛泽东选集》用字统计中，各单字用字次数之总和比《毛泽东选集》的总字数少170次；《汉字频度表》各表的总字次之和与总表的总字次竟相差31 565次之多。可见，手工查频实在是一件事倍功半，枯燥乏味的困难工作。

其实，字频统计这种十分单纯的手工作业，是特别适合于用电子计算机来做的。只要我们事先编好一个字频统计程序，然后在计算机的终端把语言资料直接键入计算机，计算机便能进行统计运算，打印出字频统计的结果。

我国用电子计算机进行汉字字频的大规模统计工作，是作为

① 贝贵琴等，《汉字频度统计——速成识读优选表》，电子工业出版社，1988年。

"现代汉语词频统计"这个国家任务的一个部分来进行的，它是"现代汉语词频统计"的第一阶段。字频统计工作由北京航空学院计算机科学工程系计算机理论教研室和国家语言文字工作委员会汉字处共同完成。他们根据抽样法的理论，将1977年至1982年出版的社会科学和自然科学文献138 000 000字的语料，抽样11 873 029字进行统计。语料来源共有四个方面：①报纸期刊，②教材，③专著，④通俗读物。抽样语料分社会科学和自然科学两大类，其中，社会科学分五个科目：

（1）社会生活　包括服装、食谱、旅游、集邮等，共抽取语料577 024个汉字，含不同汉字4 210个。

（2）人文科学　包括历史、哲学、心理学、教育学、美学、社会学等，共抽取语料1 316 964个汉字，含不同汉字5 402个。

（3）政治经济　包括财贸、统计、管理等，总共抽取语料1 644 659个汉字，含不同汉字4 889个。

（4）新闻报道　包括报纸、杂志上的各种新闻，共抽取语料1 798 467个汉字，含不同汉字4 913个。

（5）文学艺术　包括小说，散文、戏剧、说唱文学等，共抽取语料2 953 903个汉字，含不同汉字6 501个。自然科学也分成五个科目：

（1）建筑运输邮电　共抽取语料264 408个汉字，含不同汉字3 010个。

（2）农林牧副渔　共抽取语料552 761个汉字，含不同汉字3 688个。

（3）轻工业　包括电子、日用化工、塑料、食品、纺织等，共抽取语料901 003个汉字，含不同汉字4 502个。

（4）重工业　包括矿山、冶金、机械、能源等，共抽取语料68 4376个汉字，含不同汉字3 916个。

（5）基础科学：包括数学、物理、化学、生物、地理、天文等，共抽取语料1 179 764个汉字。含不同汉字4 426个。

这项汉字字频统计工作已于1985年完成，提供出13种字频统计表。其中包括：

（1）社会科学、自然科学综合字频统计表各一个。

这个统计表中，使用频率比较高的前10个汉字的有关指标如下表所示：

表1.2.1

序号	汉字	字码	画数	拼音	出现次数	出现频率	累计频率
1	的	DEHO	8	de	485786	4.0855%	4.09%
2	一	OI11	1	yī	166396	1.3994%	5.48%
3	是	VIFK	9	shì	139814	1.1758%	6.66%
4	在	ZBJN	6	zài	120984	1.0175%	7.68%
5	不	BUDD	4	bù	107418	0.9031%	8.58%
6	了	IHCLE	2	le	100708	0.8470%	9.43%
7	有	OWAR	6	yǒu	99357	0.8356%	10.26%
8	和	HEPP	8	hé	86760	0.7297%	10.99%
9	人	RNCB	2	rén	81106	0.6821%	11.68%
10	这	YEAX	7	zhè	77967	0.6557%	12.33%

由表中可看出，头10个高频汉字的累计频字已达12.33%，也就是说，平均在每100个汉字中，这10个高频汉字可出现12.33次，占了十分之一强。

（2）社会科学综合字频统计表一个。

（3）社会科学分科字频统计表五个。

（4）自然科学综合字频统计表一个。

（5）自然科学分科字频统计表五个。

以上13种字频表，分别按降频次序和汉语拼音字母顺序两种排序方法排列。

这次字频统计工作，是我国历史上利用电子计算机进行的统计规模最大、统计科目最多的一次，它不仅为现代汉字的定量研

究提供了有用的数据，而且对于汉语文教学、汉字的机械处理和信息处理的研究也有参考价值，它可以为手选照排机字盘的设计、电报码本的修订、国家标准《信息交换用汉字编码字符集·基本集》的修订以及国家标准《信息交换用汉字编码字符集·辅助集》的制定，提供有价值的事实根据。

北京语言学院在对中小学语文课本进行词频研究的同时，进行了十年制语文课文字频统计。这项统计研究，反映了十年制语文课本的用字情况，统计结果提供了一个《按出现次数多少排列的常用汉字表》，包含1 000个常用汉字，它们在520 934字的全部统计材料中，出现的总次数为409 305次，占78.57%。

《字表》中所收汉字频率最高的是“的”字，其出现次数为20 648次，出现频率为0.0396364，也就是说，平均每100个汉字中，“的”字大约要出现4次。《字表》所收汉字出现频率最低的是“悲”字，其出现次数为10次，出现频率为0.0000191。按频度高低排列的前100个汉字，在语文课本中至少都出现826次以上，总计出现次数230 946次，占统计材料的44.33%，这意味着有几乎近四成半的课文内容是用这100个汉字来表达的。《字表》中的1000常用汉字，占了中小学语文课本全部篇幅的五分之四。如果在汉语的基础教学阶段和初期学习中，挑选出这些常用汉字尽先讲授，让学生尽早掌握，将会大大加快识字教学的进度，提高语文教学的质量。

武汉大学语言自动处理研究组在RD-11微型计算机上，对著名作家老舍先生的《骆驼祥子》一书进行字频统计，计算出《骆驼祥子》全书总字数为107 360字，不同汉字数为2 413个。“的”字出现频率为4.1198%，是频率最高的字。但“他”字出现频率为2.3966%，排在第二位，与其它字表的高频汉字排列顺序不同，“他”字出现频率的提高，说明了老舍小说中常用第三人称，反映了文学作品用字的特色。另外，“车”字出现800次，“祥”字出现778次，“虎”字出现220次，“妞”字出现174次，它们的出现

频率都比较高，这与《骆驼祥子》一书的内容有关，因为这本书中常出现“拉车”、“祥子”、“虎妞”这样的字眼。对《骆驼祥子》一书的字频统计，从另一个侧面揭示了现代汉字在文学作品中使用的某些特点，它反映了作家的言语风格，是研究言语风格统计学的宝贵资料。①

为了适应语文教学、词书编纂、汉字信息处理、汉字机械处理的需要，国家语言文字工作委员会汉字处从1986年6月开始研制现代汉语常用字表，

1952年6月5日，我国教育部曾经公布过《常用字表》，收常用汉字2 000个，1964年《简化字总表》公布后，《常用字表》中的字经过精简合并，实际字数只有1 968个了。从《常用字表》公布至今30多年来，社会用字情况已发生了很大的变化，有必要重新公布一个现代汉语常用字表。

这项研究工作从已有的字典和字表中共搜集了常用字资料29种，通用字资料28种，从29种常用字资料中抽样统计15种，从28种通用字资料中抽样统计5种，共抽样统计资料20种。统计资料选定后，用计算机统计了以下内容：

（1）统计20种资料出现的不同汉字总数为8 938个。

（2）统计某个单字在20种资料中出现的次数，即确定哪些字表中收了这个汉字。例如，“的”字在统计的20种资料中都出现了，就计为20次，“牡”字只在14种资料中出现，就计为14次。

（3）在20种资料中，有的资料是根据汉字在具体文章中的出现的情况统计出来的，这样的资料叫做动态资料，有的资料则只是一般的字典或人们根据主观经验编制的字表，不能反映汉字使用的动态情况，这样的资料叫做静态资料。动态资料有6种，静态资料有14种。应该统计出现字次中静态资料内出现多少次，动

① 黄俊杰、张普、杨建清、段兴灿，《WZ-2文字处理系统对〈骆驼祥子〉进行语言自动处理》，（《语文现代化》），总第7辑，1983年，第68—84页。

表1.2.2

汉　　字	20种资料	静态资料	动态资料	平均频率
侯	17	11	6	0.0372
芬	17	11	6	0.0349
巷	17	12	5	0.0308
亦	17	11	6	0.0306
微	17	11	6	0.0287
福	17	11	6	0.0278
构	17	12	5	0.0256
尼	17	12	5	0.0228

态资料内出现多少次，并统计其平均频率。例如：

统计时除了考虑汉字的频率之外，还要考虑汉字在不同学科中的分布和使用度。

统计汉字在不同学科中的分布，可以衡量某个汉字的分布是否普遍和均匀。如果某一汉字在某一学科中的出现频率很高，在其它学科却很少出现，这说明这个汉字的分布是不均匀的。如果某一汉字不仅出现频率高，而且在多学科中都出现，这说明这个汉字的分布是均匀的。字表选字应该注意到汉字分布的均匀性。加拿大学者贝克早在1935年就提出了分布率的概念，并把这个概念应用于法语词汇的研究中。我们在研制汉字的常用字表时，应该吸取国外的这一研究成果。

1964年，尤兰德（Juilland）和洛德西盖（Chang-Rodsiguez）在计算西班牙语的词汇频率，曾经提出了使用度（usage）的公式，并用这个公式来综合地计算词的使用频率和分布情况，从而使我们对于单词在语料中的使用状况获得更客观、更准确的认识。

我国学者引用了计算词汇使用度的公式来计算汉字的使用度。计算公式如下：

$$\begin{cases} S_k = \sqrt{\sum_{i=1}^{n} (N_{ki} - N_k)^2 / n} \\ D_k = 1 - S_k / N_k \times (n-1)^{\frac{1}{2}} \\ U_k = D_k \times F_k; \quad 0 \leq D_k \leq 1 \end{cases}$$

这个公式的计算条件是假定各个分科的抽样量是均匀的。其中，N_{ki}是k号字在第i类语料中的相对频率，N_k是k号字在综合类里的相对频率，n是语料的分类数，D_k是k号字的散布系数，S_k是k号字的标准分布偏差，U_k是k号字的使用度，F_k是k号字的出现字次。

但是，在实际的汉字使用中，很难要求各个分科的抽样量保持均匀。例如，文学作品涉及的社会面最广，它的抽样量应该大一些，而有的科目涉及的社会面较窄，它的抽样量就应该小一些。因此，在计算汉字的使用度时，我国学者对上面的公式进行了调整。调整后的使用度公式是：

$$S_k = \sqrt{\sum_{i=1}^{n} (N_{ki} - N_k)^2 / n}$$

$$D_k = 1 - S_k / N_k \times (n-1)^{\frac{1}{2}}$$

$$DI_k = (L_k + 8)/18$$

$$DE_k = \begin{cases} \frac{1}{2} D_k + \frac{1}{2} DI_k & F_k \geq 0.0001 \\ DI_k & F_k < 0.0001 \end{cases}$$

$$U_k = DE_k \times F_k$$

其中，DI_k是散布系数，L_k是k号字的分布系数，DE_k是根据DI_k算出的散布系数，它取决于汉字频率的高低，当$F_k \geq 0.0001$时，得到$DE_k = \frac{1}{2}D_k + \frac{1}{2}DI_k$，它表示高频汉字的散布系数，当$F_k < 0.0001$时，得到$DE_k = DI_k$，它表示低频汉字的散布系数。

根据调整后的公式算出的使用度更为科学、更为合理。

计算汉字使用度时，语料的分类，根据北京航空学院计算机

科学工程系和国家语言文字工作委员会汉字处研究现代汉字字频统计时所用的十个科目，也就是说，语料的分类数是10。

根据调整后的使用度公式来计算现代汉字字频统计得出的有关汉字的使用度举例如下：

当然，在编制现代汉语常用字表时，应该综合地考虑各方面

表1.2.3

汉　字	出现次数	使 用 度	十科分布
侯	356	263.4	9
芬	313	246.1	9
巷	527	206.5	9
亦	1605	1274.7	10
微	3391	2547.3	10
福	2264	1915.0	10
构	6263	4261.9	10
尼	2184	1708.9	10

的因素，为此，提出了4项选字原则：

（1）根据汉字的出现频率，选取出现频率较高的字。

（2）在出现频率相同的情况下，选取学科分布广、使用度高的字。

（3）根据汉字的构词和构字能力，选取构词能力和构字能力强的字。

（4）根据汉字的实际使用情况，进一步斟酌取舍。有的字在书面语中很少使用，统计时往往统计不到，但在日常生活中却经常使用，对于这样的字，也应适当选取。

这4条原则应综合使用，不能只根据某一原则来决定取舍。

根据统计计算的结果及这4条原则，编出了《现代汉语常用字表》，共3 500字，其中常用字2 500个，次常用字1 000个。

《现代汉语常用字表》定稿后，为了检验字表中所收的常用字

是否合理，山西大学计算机科学系利用计算机抽样统计2 011 076字的语料，检测选收的常用字的使用频率。

检测结果是：

（1）2 011 076字的语料中，共有不同汉字5 141个，这5 141个汉字包含《现代汉语常用字表》中的字有3 464个，覆盖率为99.48%

（2）在3464字中，含常用字表（2 500字）中的字2 499个，覆盖率为97.97%

（3）在3 464字中，含次常用字表（1 000字）中的字965个，覆盖率为1.51%

此次检测未统计到的《现代汉语常用字表》中的37个字，基本上都是书面语中很少用到而日常生活中常用的字，是根据选字原则的第4条原则选收的。因此，通过检测，证明了《现代汉语常用字表》的收字是合理的、实用的。①

1988年3月，国家语言文字工作委员会和新闻出版署联合发布了《现代汉语通用字表》，字表共收汉字7 000个，包括《现代汉语常用字表》收入的3 500字，主要依据《印刷通用汉字字形表》，删去了其中的50字，增收854字。

制订通用字表的选材时间范围从1928年到1986年。在此时间区域内采用不等密度抽样，抽样量按时间顺序递增，以近期资料为主要的抽样对象。因社会用字与政治、经济、文化的发展有密切的关系，不同时期的用字情况不尽相同，如果只依据某一短时期的用字情况选字，则有时间的局限性。适当把统计的时间拉长，纵观各个不同时期的用字情况，就可以判断某个字的使用是否稳定。选取使用稳定的字，才能避免选字的偶然性。

通用字的选取，仍综合根据频率、使用度、构词能力以及实际使用情况等四个方面的原则来决定取舍。

---

① 傅永和，《现代汉语常用字表的研制》，（《现代汉语定量分析》），上海教育出版社，第107页。

《现代汉语通用字表》是按笔画顺序排列的。除字表正文外，在附录中，还有现代汉语通用字部首（201部）顺序表、现代汉语通用字汉语拼音字母顺序表、现代汉语通用字数据统计表。现代汉语通用字数据统计表又分为两个表：一个是信息交换用汉字编码字符集·基本集内汉字数据统计表，一个是信息交换用汉字编码字符集·基本集外汉字数据统计表。

数据统计表是按照汉字的静动态分布由多到少的顺序排列的。共分九栏。

（1）序号。

（2）汉字。

（3）静动态分布　静态资料也就是不带使用频率的字表，叫做静态字表；动态资料也就是带有使用频率的字表，叫做动态字表。共统计静态字表14个，动态字表6个，静动态字表总数共20个。汉字在静动态字表中出现的次数叫做它的静动态分布。如果某个汉字在20个静动态字表中都出现，那么，它的静动态分布就是20。

（4）静态分布　汉字在静态字表中出现的次数。

（5）平均频率　平均频率的统计根据如下五个动态字表：

——陈鹤琴1928年6月编的《语体文应用字汇》，共4261字。

——四川省教育科学院1946年8月编的《常用字选》，共2000字。

——748工程查频组1976年12月编的《汉字频度表》，共6376字。

——北京语言学院语言教学研究所1985年3月编的《汉字频率表》，共4574字。

——新华社技术研究所1987年1月编的《1986年度新闻信息流通频度》，共6001字。

——北京航空学院计算机科学工程系和国家语言文字工作委员会汉字处于1985年3月编写的《现代汉语用字频度表》，共7754字，单独列项，不参与平均频率计算。

这样，平均频率的计算公式是：

$$P = \sum_{r=1}^{5} F_r / 5$$

其中，P是平均频率，F_r是各个动态字表的频率。

（6）平均频率分布情况　数字5是指5个动态字表都统计到这个字，数字4是指4个动态字表统计到这个字，……。

（7）1985年字次　指在1985年3月编的《现代汉语用字频度表》中出现的次数，即“字次”。

（8）在《现代汉语用字频度表》中的使用度。

（9）十科分布　指的是在《现代汉语用字频度表》中分科数。数字10指的是该字出现在10个学科中，数字9指的是该字出现在9个学科中，……。

下面是这个数据统计表的头10个字的情况：

表1.2.4

序号	汉字	静动态分布	静态分布	平均频率	分布	1985年字次	使用度	十科分布
1	瓶	20	14	0.0391	5	788	556.2	10
2	熟	20	14	0.0307	5	2650	1824.8	10
3	留	20	14	0.0412	5	3615	3051.4	10
4	雨	20	11	0.0411	5	2644	2091.2	10
5	更	20	14	0.0300	5	8995	7532.1	10
6	流	20	14	0.0806	5	12369	8301.9	10
7	淡	20	11	0.0124	5	787	534.5	10
8	六	20	14	0.1350	5	9341	7772.1	10
9	语	20	11	0.0352	5	2679	2218.5	10
10	部	20	14	0.1957	5	28458	23158.1	10

为了弄清汉字在新闻信息中的流通规律，新华社技术研究所对汉字在新闻信息中的流通频率进行了统计研究。他们准备了近两年时间，设计了计算机自动统计软件，选择新华社国内通稿电

路，从1986年1月1日起到12月31日止进行统计，共统计了90 627篇稿件，汉字容量为40 632 472个。统计结果表明：1986年使用的不同汉字为6 001个，标点符号17个，外文字符39个，阿拉伯数字10个，其它字符30个，全年共使用字符6097个①。

新闻汉字流通频率的统计表明，汉字的使用带有明显的时代特征。1986年度使用频率最高的汉字依次是"的国一十中"，如果把这五个汉字的顺序重新整理一下，就是："中国的十一"，这恰恰是我国的国庆节！这种偶然的巧合，把我国人民对于自己国庆节的炽热感情表现在新闻汉字的流通使用中。"一二三四五六七八九个十百年月日"等表示数字和日期的汉字流通频率很高，反映了在改革开放的形势下，我国人民重视科学数据、重视时间和速度的特点。在各种字符的流通频率中，逗号"，"居首位，"的"字居第二位，"的"字的使用频率，从748工程《汉字频度表》中的3.75%，下降到流通频率统计时的3%(去掉标点符号所作的统计)。句长平均为每句43个汉字，段长平均为每段100个汉字，新闻每篇平均长度为401个汉字，比748工程时统计出的新闻平均长度短60%。这种情况，反映了新闻的文风逐渐简短化的趋势。748工程统计的是"文化大革命"后期的资料，当时的文章比较冗长，改革开放十年来，文章写得短小精干，文风有了明显的改进，这是令人高兴的事。

如果不统计标点符号，那么，从新闻汉字的流通频率统计中，还可以看到累计使用频率与汉字按降频顺序排列的字数之间，存在着如表1.2.5的关系。

他们再参照其它统计资料，对上述统计数字加以修正，得到表1.2.6。

根据这样的分析，他们提出，把累计使用频率在0.9—0.99之间的汉字定为常用字，累计使用频率在0.999—0.9999之间的汉字

---

① 郭冶方，《新闻信息汉字流通频度统计》，(《现代汉语定量分析》)，上海教育出版社，1989年，第95—106页。

表1.2.5

累计使用频率	字　　数	累计使用字次	最低使用字次
0.9	843	32710323	6067
0.99	2147	35978550	589
0.999	3606	36305314	57
0.9999	4872	36338004	8
0.99999	5586	36341304	2

表1.2.6

累计使用频率	字　数	重查字数	重查时最低使用字次	对字数的近似处理
0.9	843	911	6000	1200(95%)
0.99	2127	2079	600	2400(99.4%)
0.999	3606	3521	60	3600
0.9999	4372	5026	6	4800
0.99999	5586	5658	2	6000(≈100%)

定为次常用字，累计使用频率在0.99999以上的汉字定为罕用字，从而把汉字分为如下的三个区：

序号为0001—2400：常用字区

序号为2401—4800：次常用字区

序号为4801—7200：罕用区

序号为7200以上的汉字属后备字区和古汉字区。

他们认为，在次常用字区的汉字，其使用频率常因专业的不同而大起大落，频率的浮动性很强，为了照顾各专业的特点，设置次常用字是很有必要的，它可以弥补常用字区的不足。

汉语词频的统计工作比字频的统计工作更难，这是由于书面汉语不是按词分写的，而是以汉字为单位逐个连写的，词在书面上的形式不突出，因此，汉语的大规模的统计研究一直停留在以字为单位的阶段上。规模较大的汉语词频统计，仅见于台湾的刘

森（Eric Shen Liu）所编的《汉语频率词典》(1973年出版）一书，只收25000词，取样的范围和数量都非常有限。近年来，北京师范大学现代化教育技术研究所、北京语言学院语言教学研究所、北京航空学院计算机科学工程系分别进行了大规模的词频统计研究工作，目前已取得十分可喜的成果。

书面汉语是以汉字为基本字符的连续的符号串，词与词之间没有空白，在一般的情况下，我们看到的只是一个一个前后相续的汉字，而不是彼此分开的词，汉语的词被淹没在一串串没有空白的汉字流中。但是，在词频统计时，统计的基本单位是词而不是汉字，因此，必须把连续的汉字符号串按词进行切分，才能找出统计的基本单位，也才有可能进行词频统计，这种工作叫做“切词”。“切词”不仅是进行汉语词频统计的先决条件，而且，它对于汉语的计算机自动理解，对于汉外自动翻译，也都是首先必须进行的必不可少的工作。

目前切词的方式有计算机自动切词和人工切词两种。北京师范大学和北京语言学院采用人工切词的方式来进行词频统计，而北京航空学院则采用自动切词的方式来进行词频统计。

所谓“人工切词”，就是凭借人们所具有的词汇知识、语法知识以及对上下文的理解，从连续的汉字符号串中把词正确地分割出来，使词与词之间出现空白。由于参加切词的人在文化素养、专业水平方面存在差异，不同的人往往会作出不同的切分，切词的结果相差很大，就是同一个切词的人，由于记忆上的差错，前后两次切词的结果也不会完全相同。因此，在人工切词的过程中，应当相互校对，经常讨论，反复审核，把切词的误差减少到最低限度。

所谓“自动切词”，就是根据机器词典和切词原则，由计算机来切词。自动切词的算法大致可以归纳为 4 种：

（1）最大匹配法（简称MM法） 如果机器词典中最长的词为m个汉字，则取汉字字符序列前m个汉字为一个字段，查词典

匹配该字段，如果不成功，削去该字段最后一个汉字，继续查词典直到在词典中匹配上相应的单词为止。

(2) 逆向最大匹配法（简称RMM法） 其原理与 MM 法相同，但是顺序相反，从句子末尾开始，逐步由后向前推进，如匹配不成功，就去掉字段的最前面一个汉字，继续查找机器词典，直到在词典中找到相匹配的单词为止。

MM法的切词精度可以达到 1/150 字，即平均每150个汉字出现一次错误的切分，而RMM法的切词精度可以达到1/245字，即平均每245个汉字出现一次错误的切分。RMM法之所以比MM 法的切词精度高，主要在于汉语构词法的特点及方位词的影响。因此，RMM法是特别适合于汉语的切词算法。

(3) 逐词遍历匹配法：把词典中的词按照从长到短的顺序排列，逐个搜索匹配待处理的语料，直到切分出全部的单词为止。这种方法要求待处理语料中的每个切词对象都要遍历匹配机器词典中的所有单词，切词速度太慢。

(4) 最小匹配法：从一个汉字开始查机器词典进行匹配，如不成功再增加匹配长度。

从时间复杂度、空间复杂度和切词精度三方面综合考虑，一般都认为RMM法为较好的切词算法。北京航空学院采用这样的算法来进行自动切词，取得了很好的效果。

北京语言学院语言教学研究所对不同体裁和内容的200 万字(去掉标点后为181万字，共计131万个词次)的汉语语料进行了手工切分和统计，并与中国社会科学院语言研究所合作，借助于NEC-ACOS-4 电子计算机完成运算和排序。在这一工作中，他们着重地统计了全国中小学通用教材语文编写组1978—1980年编写的十年制语文课本的字频与词频。关于字频统计的情况，前面已介绍过，这里介绍他们进行词频统计的情况。

北京语言学院在对外汉语教学中，究竟应该选择哪些汉字、词语尽先教给学生？现代汉语中哪些词是常用的，哪些词是次常

用的？汉语的最低限度词汇量有多大？一个外国留学生至少要掌握多少词汇，才可以同一个中国的高中毕业生大致相当，能够适应在中国大学听课、进行讨论和书面阅读的需要？这些都是亟待解决的问题。为了在科学的基础上选择和确定现代汉语常用词语，避免汉语词汇教学的主观盲目性，提高教学效率，保证教材、辅助读物和工具书的质量，他们于1979年把“现代汉语词汇统计研究”列为重点科研项目，开始进行词频统计的研究。①

这项研究工作，采用人工与电子计算机相结合的方式，对179篇样文、近200万字的语料进行了词语切分、词频统计和数据分析工作，统计总词汇量为1315 752词次，含不同单词31 159个，其中包括十年制语文课本（52万字，374 654词次）的字频和词频的定量分析，统计结果编成《现代汉语频率词典》出版。

根据数理统计的原理，所统计材料的总体个数必需达到足够数量，才能保证统计结果符合语言的客观实际，但是，统计时又不能无限制扩大语料的范围和数量，这就产生了样本数量的最佳选取问题。词频统计属贝努利概型，可以利用相应的定理来论证取样数目的适度值。一般可以利用常用词出现频率不低于10^{-5}的先验假定（即在10万次场合，常用词至少会有一次机会出现），这时若再增大一个数量级，即选取100万字的语料，在一定意义上说就是适度的。《现代汉语频率词典》实际统计了200万字，8000个高频词出现的频率占全部语料的95％以上，每个词平均出现156次，其余23000个低频词也平均出现2.8次以上。可见，抽样总数达到130万词次，对选定日常使用的常用词来说，已经是足够大了。他们曾用随机抽样的办法，选出5万字的语料来检验频率词典中的头5000个高频词的覆盖率，结果所选5万字的语料中，有88.5％以上都出现在频率词典的头5000个高频词中，把所检验的词扩充至8000个，覆盖率达95％，可见，该频率词典的语料抽样是经济的、

① 常宝儒，《现代汉语频率词典的研制》，（《现代汉语定量分析》），上海教育出版社，1989年，第30—59页

适度的。

他们选取的语料可分为如下4类：

(1) 报刊政论 44万字，占语料总量的24.4%。

(2) 科技和科普文章 29万字，占语料总量的15.8%。

(3) 口语材料 20万字，占语料总量的11.1%。

(4) 文学作品 89万字，占语料总量的48.7%。

不过关于词频测定的最佳语料数量，目前还有不同意见。1971年英语词频统计，所用的语料量有5 088 721个词，含不同单词86 741个，其数量远比《现代汉语频率词典》的语料量大。这个问题还有待进一步的研究。

这次词频统计得出如下词表：

(1) 按字母音序排列的频率词表 该表共列出使用度大于2个的词条16593个，其中，以Z开头的词有1457个，占8.78%，以S开头的词1327个，占7.99%，以J开头的词1243个，占7.49%，以Y开头的词1205个，占7.26%，而以E和O开头的词则很少，以E开头的词64个，占0.38%，以O开头的仅13个，占0.07%。

著名语言学家赵元任曾编过一个《施氏食狮史》，文中全是以"Shi"拼音的字，致使语义的分辨发生困难，这与汉语单词中以S开头的词多不无关系。《施氏食狮史》全文如下：

> 石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，氏始试食是十狮尸。食时，始识是十狮尸，实十石狮尸。试释是事。

这篇妙趣横生的短文，靠了汉字的帮助，不难读懂，而如果有音无字，通篇的Shi Shi Shi…，恐怕是谁也听不懂的。据统计，"Shi"这个音节，在1000个音频汉字中占有24个，赵元任先生编《施氏食狮史》①，大概就是从这个事实出发的。

① 赵元任，《语言问题》，商务印书馆，1980年。

（2）按频率递降顺序排列的词表　其中，最常用词的使用频率相当高，前100个词占了语料总量的40%以上，前500个词占了语料总量的70%以上，前2562个词占了语料总量的85%，词表共有词条31159个，这些词占了语料总量的100%，从前100个词到前500个词，不同单词数增加了400个，百分比就增加了30%，而以前2562个词到31159个词，不同单词数增加了30597个，百分比才增加了15%。由此可见，高频词对于百分比的增加有着很大的作用，而低频词对于百分比的增加，其作用是很小的，往往要大量的低频词，才能使百分比增加一点点。

（3）按使用度递降顺序排列的词表　这个词表又分为两个表：使用度较高的前8 000词词表，使用度较低的词语单位表。

在使用度较高的前8 000词词表中，收入的都是使用度为6以上的词。使用度最高的词是“的”，其使用度为69 080。使用度在1 000以上的词共129个，词次累计占全部语料的44.7%；使用度在100以上的词共1230个，词次累计占全部语料的75.8%；使用度在20以上的词4 186个，词次累计占全部语料的90.1%。这说明，《现代汉语频率词典》所统计的1 314 404词次的语料中，有十分之九是用这4 000个词写成的。这些词可以成为“常用词”的候选对象。

在使用度较低的词语单位表中，收入了使用度为5及小于5的词22 446个，这些词也都是低频词，出现次数都在10次以下。如果有的词的使用度与频率比较相配，则说明这些词的分布还比较均匀，可以作为通用词的候选对象。

（4）按语体分类的高频词表　又可再分为4个表。

a.报刊政论语体的前4 000词词表：　本表共统计34种语料，29万词次（44万字），有不同词条数12 107个。前4 000个词累计频率94.77%。其中一些政治词语，如“唯心”、“党派”等，在本表中出现频率都比较高，反映了政论语体的特点。

b.科普语体中前4 000词词表　本表共统计21种语料，20万词次（29万字），有不同词条12 364个。前4 000个词累计频率92.27%。

其中，一些科技用语，如“纤维”、“合成”等，在本表中出现频率都比较高，反映了科普语体的特点。

c.生活口语中前4 000高频词词表　本表共统计18种语料，16万词次（20万字），有不同词条8 263个。前4 000个词的累计频率为96.65%。从统计数字可以看出，口语语体用词量比前两种语体要少三分之一，但高频词出现的词次却相当多，前1 000个高频词的出现频率比a表高出6%，比b表高出12%。这意味着，口语语体用词量虽然不大，但它们的出现次数对语料的覆盖面却相当大。

d.文学作品类前4 000高频词词表　本表共统计106种语料，66万词次（89万字），不同词条23 622个。前4 000个高频词累计频率为90.63%。这说明文学作品用词量大，但为了追求词汇的多样化，就是高频词使用的次数也比较低，这反映了文学作品词汇丰富多采的特点。

（5）前300个高频词分布情况分析　单词的定量分析不仅要考虑词的出现次数和频率，而且还要考虑分布情况和使用度。从理论上说，我们可以根据单词的出现次数、语料中的总词数和单词所属语料类别中的总词次，推算出有关单词的所期望具有的分布值，这种分布值叫做理想分布词次。计算公式如下：

$$\text{理想分布词次} = \text{该词总词次} \times \frac{\text{该类总词次}}{\text{语料总词数}}$$

实际分布词次与理想分布词次可能会有差异或偏离，这种差异或偏离的大小，可通过偏差系数来计算。偏差系数的计算公式如下：

$$E_k = \left(\sum_{i=1}^{4} (O_{ki} - e_{ki})^2 \Big/ e_{ki} \right) \Big/ 4$$

$$e_{ki} = e_k \times p_i$$

其中，E_k是k号词的偏差系数，e_{ki}是k号词在第i类语料里的理想分布词次，e_k为k号词在全部语料中的词次，O_{ki}是k号词在第i类语料中的实际分布词次，p_i表示第i类语料在全部语料中所

占的百分比。由于《现代汉语频率词典》的语料分为 4 大类，故以 4 类来计算偏差系数。语料中共有不同单词31 159个，故k的值可取区间1≤k≤31 159内的各整数值。

例如，"你"这个词的实际出现次数为9 694次，按使用度公式计算出的使用度是6 103，两者相差三分之一以上；按偏差系数公式计算出它的偏差系数为3466.92，在前300个高频词中，偏差系数最高。之所以出现这样大的偏离，是由于在报刊政论和科普文章中，"你"这个词用得极少，在报刊政论中，"你"只出现115次，这是它的实际分布词次，而按理想分布词次公式算出的所期望达到的理想分布词次应为2 141次，二者悬殊极大。"已经"这个词的实际出现次数、使用度在各类语料中的偏离程度都不大，按公式算出的偏差系数为1.35，在前300个高频词中，它是分布最均匀，偏差系数最小的词。

通过偏差系数来比较实际分布词次与理想分布词次之间的差异，可使我们对于前300个高频词的实际使用情况，得到更加清楚的认识。

汉语的词是由汉字组成的，因此，从《现代汉语频率词典》中，我们还可以对汉字进行定量的分析。

把频率词典中的词全部分解为汉字，共得不同汉字 4 574个，这些汉字分布于总字数为1 808 114字的语料中。其中，出现245次以上的前1 000个高频汉字，累计字次占全部语料的91.3%，如果截止到出现30次以上的前2 418个汉字，累计字次可占全部语料的99%以上，其余的出现次数低于30次的2 156个汉字，其出现字次之总和，只能占全部语料的1%，每个字的平均出现机会为千万分之五。

分析前1 000个高频汉字的语音情况，可以了解到：这1 000个汉字共以626个音节形式出现，占全部普通话语音音节（1325个）的47.1%，其中，阴平音节145个，阳平音节135个，上声音节148个，去声音节188个，轻声音节9个，无调音节1个，去声音节占绝

对优势。从组成音节的首字母来看，各首字母的音节数为：

a = 5，　b = 28，　c = 41，　d = 38，
e = 5，　f = 20，　g = 34，　h = 34，
j = 34，　k = 16，　l = 34，　m = 28，
n = 24，　p = 21，　q = 24，　r = 11，
s = 45，　t = 34，　w = 20，　x = 33，
y = 42，　z = 53。

这与单词的首字母分布情况有些接近，仍以z、s等字母开头的音节为多。

分析前1 000个高频汉字的语义情况，可以对汉字表示的基本语义项目（简称义项）得到一个较为系统化的认识。基本义项的研究可为自然语言理解和机器翻译的语义分类系统的研究提供有价值的参考。

汉字的基本义项分类及所含汉字数如下：

1. 社会生活
   1.1 社会结构 29字（国、家、县……）
   1.2 职业行业 17字（工、农、商……）
   1.3 人际关系 34字（男、女、妈……）
   1.4 人的躯体 22字（头、脑、脸……）
   1.5 食 23字（粮、食、米……）
   1.6 衣 10字（衣、服、鞋……）
   1.7 住 16字（门、窗、房……）
   1.8 行 8字（车、船、轮……）
   1.9 文化生活 20字（纸、笔、书……）
   1.10 钱物武器 16字（钱、货、枪……）
2. 自然界
   2.1 自然景观 21字（江、山、河……）
   2.2 季节时间 19字（年、月、春……）
   2.3 方位空间 30字（东、南、左……）

2.4 性状抽象 63字（形、性、意……）
2.5 矿藏资源 25字（金、铁、煤……）
3. 数质概念
3.1 数字 20字（一、千、两……）
3.2 量词 40字（种、个、次……）
4. 品质属性
4.1 颜色 8字（红、黑、白……）
4.2 一般性质 45字（古、老、美……）
4.3 对立性质 46字（大<=>小、好<=>坏、长<=>短……）
4.4 整体部分 12字（总、全、整……）
5. 行为状态
5.1 存在 15字（是、在、有……）
5.2 一般动作 23字（作、做、搞……）
5.3 社会行为 25字（改、革、调……）
5.4 外力表现 17字（映、照、燃……）
5.5 战争攻防 27字（战、乱、攻……）
5.6 始终增减 16字（始、继、减……）
5.7 人身动作
5.7.1 脸部动作 39字（看、瞧、吃……）
5.7.2 心理活动 34字（思、考、想……）
5.7.3 感情流露 14字（欢、喜、笑……）
5.7.4 手的动作 50字（拿、推、拉……）
5.7.5 腿的动作 37字（行、走、跑……）
5.7.6 全身动作 32字（飞、游、爬……）
5.8 状态 15字（站、立、坐……）
6. 指代关系 28字（我、你、他……）
7. 连续转折 20字（和、或、而……）
8. 能愿助动 14字（能、会、要……）

9. 介词　　13字（把、被、由……）

10. 副词

10.1 表示程度　15字（最、很、都……）

10.2 比较关系　7字（比、如、似……）

10.3 限定时间　12字（刚、才、就……）

11. 否定关系　6字（不、没、非……）

12. 助词　　6字（的、地、得……）

13. 语气感叹　11字（吗、呢、吧……）

前1000个高频汉字的这个基本义项分类系统，形成一个语义网络，汉语语言生活中的各个方面都大致可以通过这个语义网络来描述。

分析前1 000个高频汉字的构词能力，可以了解到，构词能力最强的汉字是"子"，它可构成单词668个，其后依次是"不、大、心、人、一、头、气、无、水"。构词能力在100条以上，出现词次在1 000次以上的前70个汉字，它们参与构成的词条达11 133条之多，占不同词条总数的35.7%。这些汉字在词首、词间、词末的构词能力也显示出明显的规律性：在词末的构词能力最强，在词首的构词能力只是在词末构词能力的87%，在词间的构词能力只是在词末构词能力的35%。从高频汉字构词能力的分布特点来看，汉语自动切词采用逆向最大匹配法自右而左地进行切分，在理论上是有根据的，因为左面词末的汉字往往就是单词的分界线，它最有可能成为自动切分的切分点。

词频统计的结果还表明：31 159个不同单词的平均词长为2.09字，也就是说，平均每个单词由2.09个汉字组成，这是静态的平均词长。如果从语料中单词出现的动态情况来计算，则可得出动态的平均词长为1.36字，也就是说，在实际的语言运用中，平均每个单词由1.36个汉字构成。

过去有许多学者认为，现代汉语以双音节词占优势，从词频统计的结果来看，这种说法未免失之片面。双音节词在31 159个

词条的词典中，所占的比例确实很大，占词条总数的73.6%，而单音节词仅占12%，三音节词占7.6%，四音节词占6.4%，五音节以上的词占0.2%。但是，在语言的使用中，在具体语料的动态环境中，双音节词只占34%，而单音节词占了64%，其余各种音节的词的词次总和不到2%。

北京师范大学现代教育技术研究所的汉语词频统计工作的研究结果，同样也以无可辩驳的统计数字，动摇了关于现代汉语中双音节词占优势的论点。他们在对106.8万字的语料统计的基础上，切词后得出了总出现次数为704 841词次，并建立了含有39 601个词的汉语频率词表。在按降频顺序排列词表的前8 000个词中，单音节词为1 413个，占17.66%，双音节词为6050个，占75.6%。但在语言使用的动态环境中，在704841个词次的语料中，单音节词出现词次为371 886次，占52.7%，双音节词出现词次为308 709，占43.8%，三音节词出现词次为18172，占2.6%，四音节词出现词次为5 929，五音节词出现词次为84，六音节词出现词次为61，这三项共占0.9%。在按降频顺序排列词表的前8 000个词中，单音节词的出现次数占前8 000个词的总出现次数的52.2%，而双音节词的出现次数仅占前8 000个词总出现次数的36.7%。可见，在语言的使用中，占优势的不是双音节词而是单音节词。

北京航空学院的现代汉语词频统计是目前国内外规模最大的汉语词频统计工作，他们在国内首次实现了现代汉语计算机自动切词，现已取得可喜的成果。

这项词频统计选材约三亿字，样本总字数达2500万字之多。他们把全部语料按时间顺序分为四个时期：

第一时期：1919—1949

第二时期：1950—1965

第三时期：1966—1976

第四时期：1977—1982

整个统计样本按学科分为社会科学和自然科学两大类，每类

又分五个子类，共十个子类。

主要成果有以下三项：

1.四个时期十类分科频率表（第一个时期没有自然科学的五类分科频率表），共35个频率表。

2.四个时期中每一个时期的社会科学综合频率表，自然科学综合频率表和社会科学、自然科学综合频率表。

3.四个时期的综合频率表。

他们的研究结果，也证实了在动态使用环境中，汉语单词以单音节词居首位。根据他们的统计，在具体的语料中，单音节词的出现词次占56.70%，双音节词的出现词次占39.65%，三音节词的出现词次占2.21%，四音节词的出现词次占1.19%，五音节词的出现词次占0.144%，六音节词的出现词次占0.083%，七音节词的出现词次占0.023%，这些统计数字也同样雄辩地证明了在现代汉语的动态使用情况下，在具体的语言中，单音节词确实是占优势的。

这次词频统计工作采用"字词混合压缩码"，可以区分多音字，使得统计结果更为精确。下面是部分多音字的频率表：

表1.2.7

汉字	读音1	出现次数	读音2	出现次数
行	xing	4543	hang	1183
重	zhong	5109	chong	1085
长	chang	8536	zhang	4934
还	huan	549	hai	28889
了	le	136183	liao	236
都	dou	37936	du	280
和	he	134281	huo	16
给	ji	17	gei	14292
省	xing	63	sheng	7059
角	jiao	4970	jue	2

统计数字向我们说明了多音字的分布情况，这对于现代汉语的规范化的研究，无疑是很有价值的。

第3节 语音统计研究

在语音自动识别与合成的系统的研制中，必须对于语音的统计特征进行研究，才有可能进一步提高系统的性能。另外，在汉字编码、文字改革的研究中，也有必要了解语言的统计特征。

中国社会科学院语言文字应用研究所拼音研究室，根据北京航空学院的现代汉字字频统计材料，在VICTOR—9000电子计算机上，对7 754个现代汉字的字音进行了统计研究，取得了现代汉字的声母、韵母、声调、音节的各种统计数据。由于这些数据是对动态使用中的汉字进行统计的结果，因此，它们更能准确地反映汉字字音在现代汉语书面语中的分布规律。①

汉字字音的统计研究分别统计了声母、韵母、声调和音节的频率。

汉字字音的声母频率统计结果如表1.3.1所示。

从表1.3.1中可以看出，22个声母中，前6个声母的出现频率可以覆盖全部声母出现频率的50%以上，前8个声母可以覆盖60%以上，前10个声母可以覆盖70%以上，前13个声母可以覆盖80%以上，前16个声母可以覆盖90%以上。

在所有声母中，零声母频率最高，占了13.9838%.除零声母外，其余的辅音声母的频率，占所有声母的86.0162%。

按辅音声母的发音部位分类，各类声母的频率分别为：

①舌尖中音（d, t, n, l） 21.9193%

① 敌小平，《现代汉语语音统计试验》，1986年。

表1.3.1

序　号	声　母	字　数	频　率(%)	累计频率(%)
1	零声母	1057	13.9838	13.9838
2	d	375	10.9514	24.9351
3	sh	332	7.4206	32.3557
4	j	618	7.1768	39.5325
5	zh	451	6.7353	46.2678
6	l	527	5.5142	51.7820
7	x	480	5.2758	57.0578
8	g	338	4.8006	61.8584
9	h	391	4.5032	66.3615
10	b	366	4.4532	70.7967
11	z	182	3.6166	74.4133
12	t	325	3.5889	78.0022
13	q	353	3.2521	81.2543
14	ch	328	3.0355	84.2898
15	m	320	3.0184	87.3082
16	f	233	2.7952	90.1034
17	r	106	2.0438	92.1472
18	n	167	1.8648	94.0120
19	s	183	1.8346	95.8644
20	k	222	1.7289	97.5755
21	c	154	1.3071	98.8826
22	p	243	1.1174	100.0000
合　计		7754	100.0000	

② 舌尖后音（zh, ch, sh, r）　19.2352%

③ 舌面音（j, q, x）　15.7047%

④ 舌根音（g, k, h）　11.0327%

⑤ 双唇音（b, P, m）　8.5710%

⑥ 舌尖前音（z，c，s） 6.7583%

⑦ 唇齿音（f） 2.7952%

按辅音声母的发音方法分类，各类声母的频率分别为：

① 塞音（b，P，d，t，g，k） 26.6224%

② 塞擦音（z，c，zh，ch，j，q） 25.1234%

③ 擦音（f，s，sh，r，x，h） 23.8732%

④ 边音（l） 5.5142%

⑤ 鼻音（m，n） 4.8832%

在塞音与塞擦音中，送气音（P，t，k，c，ch，q）频率为14.0299%，不送气音（b，d，g，z，zh，j）频率为37.7159%。

按辅音声母的清浊分类，各类声母的频率分别为：

① 浊音声母（m，n，l，r） 12.4412%

② 清音声母（即其它辅音声母） 73.5750%

汉字字音的韵母频率统计结果如下：

表1.3.2

序号	韵母	字数	频率(%)	累计频率(%)
1	i	1092	16.4993	16.4993
2	e	243	10.2218	26.7211
3	u	652	6.7100	33.4311
4	ian	376	4.3407	37.7717
5	ong	203	3.9421	41.7138
6	uo	232	3.7303	45.4441
7	ing	263	3.6484	49.0925
8	ai	215	3.5357	52.6281
9	an	423	3.4684	56.0699
10	a	255	3.4319	59.5285
11	eng	222	3.1599	62.6884
12	en	181	3.0292	65.7176
13	uei	274	3.0024	68.7199

续表

序　　号	韵　　母	字　　数	频　率(%)	累计频率(%)
14	ao	302	2.7597	71.4797
15	ang	246	2.6626	74.1423
16	iou	158	2.5771	76.7194
17	iao	262	2.2610	78.9803
18	ü	275	2.2457	81.2260
19	ie	209	2.1406	83.3666
20	in	197	2.0805	85.4471
21	iang	126	2.0613	87.5085
22	ou	193	1.8418	89.3503
23	uan	161	1.5063	90.8565
24	ei	133	1.3084	92.1649
25	ia	96	1.1232	93.2882
26	üan	101	0.9996	94.2878
27	uen	129	0.9525	95.2403
28	üe	82	0.9435	96.1838
29	uang	94	0.7962	96.9800
30	ua	53	0.6216	97.6016
31	er	13	0.6183	98.2201
32	iong	51	0.5483	98.7684
33	ün	86	0.5130	99.2814
34	o	106	0.3674	99.6847
35	uai	40	0.3465	99.9953
36	ueng	7	0.0028	99.9980
37	io	2	0.0019	100.0000
38	m	1	0.0000	100.0000
合　　计		7754	100.0000	

另外，ŋ，ê可以自成音节，亦可算为韵母，这样，现代汉字

的韵母共40个，但因ng，ê在所统计的7754个汉字中未出现，故在表1.3.2中未计入。

从表1.3.2中可以看出，40个韵母中，前7个韵母可覆盖全部韵母的50%，前10个韵母可覆盖近60%，前14个韵母可覆盖70%以上，前18个韵母可覆盖80%以上，前23个韵母可覆盖90%以上。

在所有韵母中，i的出现频率最高，占了16.4993%，韵母i实际上包含三个音素：一个是舌面元音[i]，一个是舌尖前元音[ɿ]，一个是舌尖后元音[ʅ]。[i]的频率是9.4332%，[ɿ]的频率是1.4946%，[ʅ]的频率是5.5715%。

韵母一般可分为单韵母、鼻韵母、复韵母三种，其频率分别为：

① 单韵母 43.4174%

② 鼻韵母 33.7118%

③ 复韵母 26.1936%

韵母按四呼可分为开口呼、齐齿呼、合口呼、撮口呼四种，其频率分别为：

① 开口呼韵母 43.4714%

② 齐齿呼韵母 29.6660%

③ 合口呼韵母 21.6107%

④ 撮口呼韵母 4.7018%

韵母按韵头的有无，又可分为无韵头韵母和有韵头韵母两种，其频率分别为：

① 无韵头韵母 72.5944%

② 有韵头韵母 27.4056%

在有韵头韵母中，齐齿呼韵头韵母的频率为14.5039%，合口呼韵头韵母的频率为 10.9586%，撮口呼韵头韵母的频率为 1.9431%。

韵母按韵尾的有无，又可分为无韵尾韵母和有韵尾韵母两种，

其频率分别为：

①无韵尾韵母　　　　　　　　48.6556%

②有韵尾韵母　　　　　　　　51.3444%

在有韵尾韵母中，元音韵尾为-i的韵母频率为8.1930%，元音韵尾为-u的韵母频率为 9.4396%，故有元音韵尾的韵母频率共为17.6326%；鼻音韵尾为-n的韵母频率为16.8902%，鼻音韵尾为-ŋ的韵母频率为16.8216%，故有鼻音韵尾的韵母频率为33.7118%。

汉字字音的声调频率统计结果如下：

表1.3.3

序号	声调	字数	频率(%)	累计频率(%)
1	去声	2438	35.7254	35.7254
2	阳平	2016	20.5069	56.2323
3	阴平	1927	20.4313	76.6635
4	上声	1304	17.3845	94.0120
5	轻声	39	5.9880	100.0000
合计		7734	100.0000	

这样的声调频率是根据在动态使用中的汉字统计出来的。我国学者根据字表和词表中的汉字的静态统计结果，也是以去声字最多，这种情况，与我们的经验是吻合的。

再谈谈汉字字音的音节频率统计结果。

统计时分不带调音节和带调音节两种情况来进行。如果统计时不考虑音节的声调，则得出不带调音节419个，如果统计时考虑音节的声调，则得出带调音节1 333个。

不带调音节419个中，有13个音节有音无字，前20个频率最高的不带调音节如表1.3.4：

表1.3.4

序号	音节	字数	频率(%)	累计频率(%)
1	de	7	4.4762	4.4762
2	shi	82	3.4760	7.9522
3	yi	133	2.8273	10.7794
4	zhi	96	1.7608	12.5403
5	ji	127	1.6910	14.2313
6	you	45	1.3999	15.6311
7	bu	23	1.3076	16.9387
8	li	82	1.3009	18.2396
9	zhong	18	1.1845	19.4240
10	zai	10	1.1383	20.5624
11	wei	67	1.0949	21.6573
12	he	39	1.0824	22.7397
13	zhe	24	1.0657	23.8053
14	guo	18	0.9857	24.7911
15	qi	80	0.9832	25.7743
16	ge	34	0.9546	26.7289
17	ren	19	0.8735	27.6024
18	yü	106	0.8716	28.4740
19	le	9	0.8699	29.3439
20	jian	73	0.8115	30.1554

从表1.3.4中可看出，同一音节所含字数最多的是133个(yi音节)。

带调音节1 333个中，有92个音节有音无字，前20个频率最高的带调音节如表1.3.5所示。

从表1.3.5中可以看出，同一音节所含字数最多的为73个(yì音节)。

表1.3.5

序 号	音 节	字 数	频 率(%)	累计频率(%)
1	de(轻声)	3	4.1755	4.1755
2	shì	40	1.9279	6.1035
3	yī	14	1.4863	7.5898
4	bù	11	1.2740	8.8637
5	zaì	3	1.1223	9.9861
6	shí	17	1.0846	11.0706
7	hé	26	1.0474	12.1180
8	yoǔ	8	0.8640	12.9820
9	le(轻声)	2	0.8482	13.8301
10	weì	21	0.8109	14.6410
11	tā	9	0.7518	15.3928
12	gè	5	0.7123	16.1051
13	zhì	47	0.6954	16.8005
14	rén	3	0.6899	17.4905
15	lì	42	0.6826	18.1730
16	yǐ	15	0.6661	18.8391
17	zhè	5	0.6646	19.5038
18	zhōng	9	0.6565	20.1602
19	dì	20	0.6202	20.7804
20	yì	73	0.5906	21.3711

如果不从声、韵、调的角度来考虑，而把汉字转写为汉语拼音的话，那么，从动态使用中统计出的汉语拼音字母的频率如表1.3.6。

为拼写11 873 453个汉字，共需使用35 706 037个字母，平均字长为3.0072个字母。如果用数字标调法，分别用1，2，3，4，5代表阴平、阳平、上声、去声、轻声，则拼写11 873 453个汉字共需要使用46 808 513个符号，平均每个汉字使用3.9473个符号。

表1.3.6

序号	字母	频率(%)	累计频率(%)	序号	字母	频率(%)	累计频率(%)
1	i	13.7557	13.7557	15	x	1.7544	89.8513
2	n	11.8304	25.5861	16	b	1.4748	91.3262
3	a	9.9476	35.5337	17	c	1.4441	92.7702
4	u	7.7564	43.2901	18	t	1.1934	93.9637
5	e	7.6074	50.8975	19	w	1.1004	95.0641
6	h	7.2142	58.1116	20	q	1.0814	96.1544
7	g	7.1901	65.3017	21	m	1.0037	97.1492
8	o	5.4120	70.7138	22	f	0.9295	98.0787
9	d	3.6417	74.3554	23	r	0.8853	98.9640
10	z	3.2504	77.6059	24	k	0.5749	99.5390
11	y	3.1932	80.7991	25	p	0.3716	99.9105
12	s	3.0777	83.8767	26	ü	0.0895	100.0000
13	j	2.3865	86.2633	合计		100.0000	
14	l	1.8337	88.0969				

对同音字调查的结果如表1.3.7（调查时按带调音节来决定同音与否，也就是说，要考虑音节的声调的异同）。

从表1.3.7可看出，多数音节含同音字数并不多，含同音字不超过3个的音节有557个，占有字音节的45%，累计频率占25.5%；含同音字不超过10个的音节有1029个，占有字音节总数的83%，累计频率占60.6%；而同音字超过20个音节只有39个，占有字音节总数的3%，总出现频率也只占13%，通过这些统计数字，我们可以对于汉语中的同音字问题，获得更具体、更准确的认识。

汉语是声调语言，声调是汉语语音系统的重要组成部分。汉语声调的类型、连续变调、辨义功能以及它与语调、语境的关系是十分复杂的，因此，有必要在定性描述的同时，对汉语声调的

表1.3.7

含同音字数	音节数	频率(%)	累计频率(%)
0(无字音节)	92	0.0000	0.0000
1(单音字)	258	6.4037	6.4037
2	156	4.9301	11.3441
3	143	14.1586	25.4927
4	93	4.2937	29.7864
5	88	6.1753	35.9617
6	73	5.1064	41.0681
7	63	4.8003	45.8684
8	58	5.6494	51.5178
9	45	5.0619	56.5797
10	44	3.9847	60.5646
11	27	3.4594	64.0240
12	28	2.3847	66.4086
13	22	2.7503	69.1589
14	23	3.6570	72.8159
15	16	2.5488	75.3647
16	17	2.6790	78.0437
17	14	2.8476	80.8913
18	11	2.6485	83.5838
19	4	0.4074	83.9472
20	11	3.0315	86.9787
21	7	2.0626	89.0414
22	2	0.3366	89.3779
23	3	0.4787	89.8566
24	3	0.6251	90.4817
25	2	0.1786	90.6603
26	3	1.8538	92.5141
27	0	0.0000	92.5141
28	3	0.3254	92.8395

类型、分布、结构进行定量的统计分析。

中国社会科学院语言文字应用研究所在35220条双音节词、5423条三音节词、4354条四音节词共44997条词的现代汉语普通话词汇数据库范围内，对普通话的声调进行了一些静态统计，以便从统计学意义上揭示汉语声调的统计规律性[①]。在多音节词或词组中四声、轻声的搭配组合，叫做"声调结构"。声调结构的频率，可以反映普通话四声、轻声组合时表现出的统计规律。

双音节词声调结构频率如下表所示：

表1.3.8

序号	声调结构	出现次数	频率(%)	累计频率(%)	例词
1	去 去	4085	11.59	11.59	案 件
2	阴 去	3182	9.03	20.62	安 定
3	阳 去	2988	8.48	29.10	长 度
4	阴 阳	2370	6.72	35.82	编 辑
5	上 去	2328	6.60	42.42	保 证
6	阳 阳	2312	6.56	48.98	才 华
7	去 阳	2302	6.53	55.51	必 然
8	阴 阴	2164	6.14	61.65	参 观
9	去 上	1750	4.96	66.61	部 首
10	去 阴	1724	4.89	71.50	办 公

双音节词声调结构有20种，表1.3.8中只列出了前10种，序号1—3的三种声调结构所对应的词汇数占整个双音节词的29.1%，序号1—7的7种声调结构所对应的词汇数占整个双音节词的55.51%，其中，出现频率最高的是"去声＋去声"结构，占11.59%。

三音节词声调结构频率如表1.3.9所示。

三音节词共有声调结构100种，每种结构都存在着对应的普通话词汇，表1.3.9中只列出了出现频率较高的前10种。出现频率最

① 刘连元，马亦凡，《普通话声调分布和声调结构频度》，《语文建设》，1986年，第3期。

高的声调结构是"去声+去声+去声",占全部三音节词的3.42%。

表1.3.9

序号	声调结构	出现次数	频率(%)	累计频率	例词
1	去去去	186	3.42	3.42	备忘录
2	阳去去	139	2.56	5.98	门外汉
3	去阳去	126	2.32	8.30	半瓶醋
4	阴阳去	122	2.24	10.54	风凉话
5	去阴阴	120	2.21	12.75	未婚妻
6	去去阳	115	2.12	14.87	大概其
7	阴去去	113	2.08	16.95	方块字
8	去阳阳	111	2.04	18.99	乱弹琴
9	阴阴去	107	1.97	20.96	单身汉
10	去阴去	102	1.88	22.84	办公室

四音节词声调结构频率如下表所示:

表1.3.10

序号	声调结构	出现次数	频率(%)	累计频率(%)	例词
1	阴阴去去	104	2.38	2.38	方兴未艾
2	阴阳去去	82	1.88	4.26	风平浪静
3	阳阳去去	69	1.58	5.84	惩前毖后
4	阳阴去去	57	1.30	7.14	防微杜渐
5	去去去去	49	1.12	8.26	背信弃义
6	去阳去去	47	1.07	9.33	不伦不类
7	去阴去去	45	1.03	10.36	幸灾乐祸
8	阴阴上去	42	0.96	11.32	颠三倒四
9	阴阳上去	42	0.96	12.28	兵强马壮
10	阴阳去上	42	0.96	13.24	班门弄斧

四音节词可能的声调结构共500种,在数据库所存词汇范围内存在对应普通话词汇的声调结构339种,占67.8%;不存在对应

普通话词汇的声调结构161种，占32.2%。没有普通话词汇的声调结构都是包含一个或多个轻声音节的，其中只有一种例外，是"去声+上声+上声+上声"。表1.3.10中只列出了出现频率较高的前10种，出现频率最高的四音节词声调结构是"阴平+阴+去声+去声"，占2.38%。

表1.3.11是双、三、四音节声调结构按频率递降顺序排列的分布数据，频率范围分为四段，每段覆盖大约四分之一的频率范围，从表中可以看出不同声调结构出现的频繁程度，也可以看出哪些声调结构覆盖了多大的频率范围。

表1.3.11

		第1段	第2段	第3段	第4段
双音节词	序号	1—3	4—7	8—11	12—20
	声调结构数目	3	4	4	9
	百分比	29.1%	26.41%	20.56%	23.79%
	累计	29.1%	55.51%	76.07%	99.86%
三音节词	序号	1—12	13—28	29—49	5—100
	声调结构数目	12	16	21	51
	百分比	26.48%	24.06%	24.47%	24.51%
	累计	26.48%	50.54%	75.01%	99.52%
四音节词	序号	1—24	25—66	67—133	134—500
	声调结构数目	24	42	67	367
	百分比	25.12%	25.23%	24.88%	23.14%
	累计	25.12%	50.35%	75.23%	98.37%

前面说过，根据在动态使用中的汉字进行统计，普通话中以去声最多，占了动态使用中的普通话全部声调的35.7254%，那么，普通话声调在静态的词汇数据库中的分布情况又怎样呢？

在词汇数据库中存储的44 977条词汇中，音节总数为104 123个，声调统计结果如表1.3.12：

表1.3.12

序号	声调	出现次数	频率(%)	累计频率(%)
1	去声	33 560	32.2311	32.2311
2	阳平	25 130	24.1349	56.3660
3	阴平	24 690	23.7123	80.0783
4	上声	17 853	17.1461	97.2244
5	轻声	2 890	2.7756	100.0000
合计		104 123	100.0000	

表1.3.12与表1.3.3相比较，可以看出，不论在动态使用中还是在静态词汇数据库中，普通话的声调按频率递减的顺序都是：去声→阳平→阴平→上声→轻声。

如果把上面的结果再与国家标准汉字编码字符集中所收的6 763个汉字的声调比较，可以发现这样的声调分布顺序仍然不变。

在6 763个汉字中，如果把一字多调和一字多音看成是不同的字，那么，这6 763个汉字就变为了7 778个字，多出1015个字。表1.3.13中列出了7 778个字中的声调分布情况。

表1.3.13

序号	声调	字数	频率(%)	累计频率(%)
1	去声	2489	32.0005	32.0005
2	阳平	1972	25.3536	57.3541
3	阴平	1959	25.1864	82.5405
4	上声	1300	16.7138	99.2543
5	轻声	58	0.7457	100.0000
合计		7778	100.0000	

上面讲的声调分布是就一个个孤立的音节来看的声调分布，我们还有必要研究在多音节的单词和词组中声调在单词和词组的不同位置分布的情况，这样的研究有助于我们了解单词和词组的

语音结构的规律性。统计结果表明，普通话的声调在词汇中的分布呈现出有趣的统计规律性。下面分别加以说明。

词汇数据库中的双音节词总数为35 220条，声调在双音节词的首音节和末音节的分布数据如表1.3.14所示。

表1.3.14

	首音节		末音节	
	出现次数	百分比(%)	出现次数	百分比（%）
阴平	9910	60.33	6516	39.67
阳平	8624	50.59	8424	49.41
上声	6283	52.16	5763	47.84
去声	10404	45.28	12575	54.72
轻声	0	0	1950	100.00

从表1.3.14中可看出，双音节词中阴平在首音节居多，去声在末音节居多；阳平和上声的分布情形差不多，都是从首音节到末音节略有减小，呈近似均匀分布。

词汇库中的四音节词总数为4 354条，声调在四音节词的首音节、第二音节、第三音节和末音节的分布数据如表1.3.15所示。

表1.3.15

	首音节		第二音节		第三音节		末音节	
	出现次数	百分比%	出现次数	百分比%	出现次数	百分比%	出现次数	百分比%
阴平	1263	29.54	1123	26.27	1030	24.09	859	20.09
阳平	1129	26.32	1145	26.69	1086	25.31	930	21.68
上声	746	25.25	755	25.55	721	24.40	733	24.81
去声	1210	21.16	1244	21.76	1504	26.31	1759	30.77
轻声	0	0	85	49.71	13	7.60	73	42.69

从表1.3.15中可看出，声调在四音节词中的分布与在双音节词中的分布有相同特点，都是阴平多居首音节，去声多居末音节。

为了更清楚地比较声调在双音节词和四音节词中的分布特点，根据统计数字绘成如下的直方图（图1.3.1）：

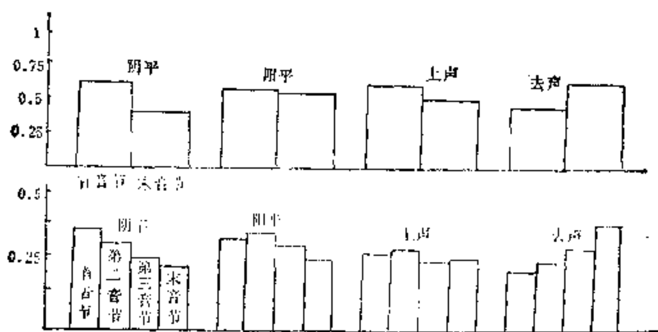


图1.3.1 声调分布的直方图

从图1.3.1中可以看出，普通话声调在双音节词和四音节词中的分布具有共同的统计规律：阴平首音节居多，从首音节到末音节阴平分布数据是递减的；去声末音节居多，从首音节到末音节去声分布数据是递增的；阳平和上声从首音节到末音节呈现略降趋势，近似均匀分布。阴平的调型是高平，去声的调型是全降。发高平调时，声带振动频率高，声门保持紧张状态；发全降调时，声带振动频率逐渐下降，声门由紧张变松弛。高平和全降在首末音节中的统计分布规律似乎反映了声带在发一个单词或词组过程中，先紧后松、由紧变松的总的自然态势。

词汇数据库中三音节词总数是5 423条，声调在三音节词的首音节、第二音节和末音节的分布数据如表1.3.16所示。

从表1.3.16中可以看出，声调在三音节词中的分布还保留着它们在双音节词和四音节词中分布的基本特点：阴平、阳平、上声从首音节向末音节的分布是减少的；去声从首音节向末音节的分布是增加的。不同的是分布的上升和下降的幅度变化比较平缓，大多数呈现出向均匀分布过渡的趋势，而且，阳平在第二音节时

表1.3.16

	首音节		第二音节		末音节	
	出现次数	百分比%	出现次数	百分比%	出现次数	百分比%
阴平	1435	35.97	1399	35.07	1155	28.95
阳平	1264	33.29	1324	34.87	1209	31.84
上声	1078	37.80	890	31.21	884	31.00
去声	1647	33.86	1525	31.35	1692	34.79
轻声	0	0	286	37.19	483	62.81

稍有升高。三音节词的构成方式一般有三种情况：

①单音节＋双音节（如“乱弹琴”）

②单音节＋单音节＋单音节（如“实打实”）

③双音节＋单音节（如“办公室”）

这三种构成方式总体来说是左右对称的，这种对称性可能是导致声调在三音节词中向均匀分布过渡的原因。

从表1.3.14、表1.3.15和表1.3.16中，我们还可以看到普通话中轻声分布的特点：双、三、四音节词的首音节上都不出现轻声；双音节词轻声都在末音节上；三音节词轻声也集中在末音节上，占62.81%；四音节词轻声集中在第二、第四音节上，分别占49.71%和42.69%。有趣的是，轻声在四音节词中的统计分布与音乐中每小节4拍“强—弱—次强—弱”的格式十分类似，这种轻声与非轻声音节的交错配合，使得四音节词的节奏显得很分明，这是汉语语音富于音乐性在统计方面的一个明证。

第4节　方言研究中的统计方法

我国方言复杂，语言学者很早就进行了方言的调查工作，积累了大量的资料，取得了很大的成绩。近年来，我国学者把数学

方法引入了方言的研究，对汉语方言进行分区的数量测定，这是现代汉语诸要素定量分析的一项新的探索①。

对汉语方言进行分区的统计方法，就是通过对一系列较能说明方言差异的项目或特征进行统计，把方言之间的异同综合成数量指标，然后用这些数量指标去找出方言分区的条理来。

汉语方言的复杂性在世界语言中是屈指可数的，各方言都有自己的特点，它们跟普通话比较，存在着程度不同的联系和差别，各大方言之间、方言区内部各次方言之间，也存在着不同程度的联系和差别，由于汉语方言间的差别突出地表现在语音方面，所以，可以根据对语音的有关数据进行数量分析，从而来判断方言之间的亲疏关系，找出对方言进行分区的理论依据。

调查点共选取了如下17个：北京、济南、西安、太原、汉口、成都、扬州、苏州、温州、长沙、双峰、南昌、梅县、广州、厦门、潮州、福州，在北京大学中文系编写的《汉语方音字汇》中，对这17个方言点的2700多字都分别注有以切韵、等韵及韵图为依据的中古音和用国际音标标明的现代读音，它们基本上能反映这17个方言点的语音面貌。将《汉语方音字汇》输入计算机、建立了机读的电子字典DOC，这样，便可直接利用DOC来进行统计和分析。

汉语的音节是由声母、韵母和声调组成的。声母、韵母处于音段层次，声调处于超音段层次，音节则是音段层次和超音段层次的统一体。因此，对方言读音的比较可以分两个方面来进行，一个方面是声母和韵母，另一个方面是声调。

声母中古声纽分为40单元："帮、滂、并、明、非、敷、奉、微、端、透、定、泥、知、彻、澄、见、溪、群、疑、精、清、从、心、邪、庄、初、崇、生、章、昌、船、书、禅、影、晓、匣、云、以、来、日"。每一个单元都排比了中古同一声母的字在

① 陆致极，《汉语方言定量分析的理论模型》，(《现代汉语定量分析》)，上海教育出版社，1989年。

现代17个方言点中声母分布的数据。例如，“帮”母单元的92字，在现代方言中声母分化为〔p、p'、b、m〕，它们的分布状况如下：

表1.4.1　　中古“帮”母在现代方言中的分布

声母＼分布数＼方言点	北京	济南	西安	太原	汉口	成都	扬州	苏州	温州	长沙	双峰	南昌	梅县	广州	厦门	潮州	福州
[p]	88	89	88	89	88	87	89	90	89	88	85	86	85	87	85	86	85
[p']	3	2	3	2	3	5	3	0	0	4	6	6	7	5	7	6	7
[b]	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
[m]	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

韵母根据中古韵摄归为16类：“通、江、止、遇、蟹、臻、山、效、果、假、宕、梗、曾、流、咸、深”。每一类又根据中古韵部、四声和“等呼”的情况分为若干单元。例如，“通”摄类的186字可以分成如下8个单元：

表1.4.2　　“通”摄类的8个单元

单元	韵　部	等　呼	字数	单元	韵　部	等　呼	字数
1	东董送	合口一等	48	5	冬　宋	合口一等	6
2	东董送	合口三等	27	6	沃	合口一等	1
3	屋	合口一等	16	7	钟肿用	合口三等	45
4	屋	合口三等	25	8	烛	合口三等	18

“通”摄单元1中属中古“东、董、送”韵合口一等的48字，在现代17个方言点中韵母分布状况如表1.4.3所示。

声调的情况比值复杂。中古音系声调的调类有平、上、去、入四类，它们在现代方言中演变分合的情况与中古声母的清和浊，送气和不送气有着密切关系，所以，把它们按表1.4.4归类。

中古平声清声母（如“春天”）的567字，在现代17个方言点中的分布情况如表1.4.5所示。

得出了声母韵母的数据和声调的数据之后，利用计算机对这

表1.4.3　　中古“东董送”韵合口一等字在现代方言中分布

方言点 / 分布数 / 韵母	北京	济南	西安	太原	汉口	成都	扬州	苏州	温州	长沙	双峰	南昌	梅县	广州	厦门	潮州	福州
[uŋ]	42	42	0	45	8	0	0	0	0	0	0	48	48	48	0	0	23
[oŋ]	0	0	43	0	40	48	0	48	48	48	0	0	0	0	0	28	0
[ɔuŋ]	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0	0	0
[aŋ]	1	0	0	0	0	0	0	0	0	0	47	0	0	0	10	19	0
[ɔŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
[ɸyŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
[əyŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
[əŋ]	3	4	3	3	0	0	0	0	0	0	1	0	0	0	0	0	0
[ouŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
[uəŋ]	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[uoŋ]	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[iã]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
[iaŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
[ɪŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
[eŋ]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

表1.4.4　　中古声调的分类

古调类	平			上			去			入		
古声母	清	全浊	次浊	清	次浊	全浊	次清	全清	浊	清	次浊	全浊
例字	春天	群	名	短体	老卵	巨抢	快	半	共洞	八德	目日	白达

些数据进行相关系数的统计，得到每两个方言点之间在声母韵母方面和声调方面的相关系数值。这些相关系数值表示了方言点之间相互关系的密切程度（见表1.4.6和表1.4.7）

从表1.4.6中可以看出在声母韵母方面各方言点之间的接近程度。例如，北京话与西安话之间的相关系数是0.8847，与苏州

表1.4.5　　中古平声清声母字在现代方言中的分布

调类＼分布数＼方言点	北京	济南	西安	太原	汉口	成都	扬州	苏州	温州	长沙	双峰	南昌	梅县	广州	厦门	潮州	福州
平声	0	0	0	556	0	0	0	0	0	0	0	0	0	0	0	0	0
阴平	564	535	554	0	549	549	547	546	547	541	534	536	538	550	529	530	545
阳平	2	6	4	0	9	6	6	3	3	7	10	2	10	8	12	8	4
上声	1	3	5	7	4	3	7	5	0	6	4	6	5	0	4	0	6
阴上	0	0	0	0	0	0	0	0	3	0	0	0	0	2	0	7	0
阳上	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
去声	0	3	4	1	5	6	2	0	0	0	0	0	10	0	0	0	0
阴去	0	0	0	0	0	0	0	4	6	13	11	7	0	4	11	9	8
阳去	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	1	0
入声	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
阴入	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

话是0.5493，与广州话是0.4330，可见，北京话与西安话之间的接近程度远甚于它与苏州话、广州话的接近程度，而北京话与苏州话的接近程度又超过了它与广州话的接近程度。

从表1.4.7中可以看出在声调方面各方言点之间的接近程度。例如，北京话与西安话之间的相关系数是0.9827，与苏州话是0.7023，与太原话是0.3076，可见，在声调方面，北京话与西安话之间的接近程度远甚于它与苏州话、太原话的接近程度，而北京话与苏州话的接近程度又超过了它与太原话的接近程度。太原话与各个方言点在声调方面都有很大的距离，与它最接近的是梅县话，与它相差最大的是温州话，正是由于太原话在声调方面的显著特点，以太原话为代表的山西方言很早就引起了语言学家的注意。

为了求得在语言上对汉语各方言点的相互关系有一个总的认识，需要把声母韵母和声调两方面的因素综合起来考虑。因此，把每两个方言点间在声母韵母上和在声调上的相关系数相加再求

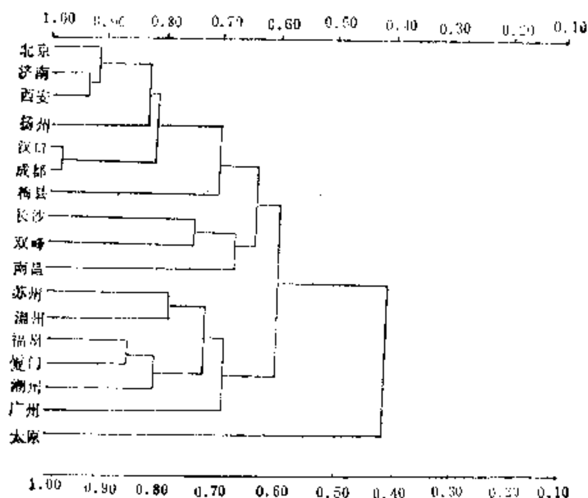


图1.4.1　汉语方言亲疏关系（声韵调）聚类树形图

表1.4.6　各方言点间的声母

	北京	济南	西安	太原	汉口	成都	扬州	苏州
北京	1.0000							
济南	·8872	1.0000						
西安	·8847	·8989	1.0000					
太原	·8003	·7541	·8155	1.0000				
汉口	·7497	·6407	·7412	·7325	1.0000			
成都	·7278	·6234	·7147	·7183	·9452	1.0000		
扬州	·7163	·7479	·7385	·7741	·6682	·6754	1.0000	
苏州	·5493	·5328	·5859	·5831	·5300	·5352	·5980	1.0000
温州	·5153	·5128	·5553	·5613	·4614	·4681	·5674	·7260
长沙	·6823	·5978	·6481	·5933	·8026	·7885	·5718	·5047
双峰	·5129	·5047	·5194	·4450	·6478	·6297	·4873	·5409
南昌	·6916	·6371	·6826	·7437	·6499	·6626	·7060	·5967
梅县	·5688	·5119	·5689	·6379	·5368	·5448	·5827	·5736
广州	·4380	·4344	·4414	·4390	·3459	·3176	·4317	·4038
厦门	·5439	·4950	·5436	·5645	·5142	·5097	·5736	·5779
潮州	·5800	·5470	·6210	·6331	·5727	·5656	·6120	·6306
福州	·5950	·5989	·6133	·6580	·5656	·5433	·6571	·5816

韵母相关系数表

温州	长沙	双峰	南昌	梅县	广州	厦门	潮州	福州
1.0000								
•4448	1.0000							
•4387	•6010	1.0000						
•5675	•5480	•4781	1.0000					
•5568	•5029	•3847	•8039	1.0000				
•4173	•3360	•2990	•5137	•6221	1.0000			
•5353	•5217	•4137	•6610	•7391	•5495	1.0000		
•5844	•5507	•4544	•7008	•7654	•5745	•8573	1.0000	
•5600	•5698	•4450	•6340	•6753	•5453	•7362	•7856	1.0000

表1.4.7　　各 方 言 点 间 的 声 调

	北京	济南	西安	太原	汉口	成都	扬州	苏州
北京	1.0000							
济南	.9931	1.0000						
西安	.9827	.9926	1.0000					
太原	.3076	.3008	.2988	1.0000				
汉口	.9670	.9480	.9441	.2988	1.0000			
成都	.9654	.9459	.9421	.3061	.9999	1.0000		
扬州	.9123	.9312	.9275	.3018	.9222	.9214	1.0000	
苏州	.7023	.6931	.6936	.1142	.6868	.6863	.6838	1.0000
温州	.5693	.5642	.5631	.0106	.5571	.5551	.5521	.8608
长沙	.7225	.7119	.7136	.0656	.7068	.7062	.7869	.8825
双峰	.7326	.7139	.7151	.0634	.7664	.7666	.6980	.9141
南昌	.6229	.6145	.6144	.0665	.6096	.6094	.6919	.8083
梅县	.9381	.9283	.9246	.3417	.9185	.9177	.9160	.7797
广州	.5741	.5692	.5680	.0398	.5620	.5601	.5572	.8485
厦门	.7053	.6952	.6967	.1268	.6901	.6895	.6871	.9918
潮州	.5875	.5824	.5813	.0142	.5748	.5728	.5698	.8219
福州	.7070	.6970	.6985	.1254	.6916	.6909	.6886	.9910

相关系数表

温州	长沙	双峰	南昌	梅县	广州	厦门	潮州	福州
1.0000								
.7459	1.0000							
.7748	.9078	1.0000						
.6677	.8864	.8243	1.0000					
.6517	.7137	.7082	.6126	1.0000				
.9794	.7608	.7936	.6831	.6262	1.0000			
.8539	.8960	.9246	.8193	.7756	.8422	1.0000		
.9589	.7337	.7419	.6284	.6706	.9329	.8158	1.0000	
.8573	.8959	.9241	.8207	.7761	.8455	.9994	.8183	1.0000

平均值，然后对这样的数值矩阵作聚类分析程序的运算，就可得到汉语方言亲疏关系（声韵调）聚类树形图。树形图的横座标是相似性尺度，标尺从1到0，表示平均接近程度的递减或平均差异程度的递增。

从图1.4.1的聚类树形图中可以看出,太原话与各方言的接近程度都比较差,独立出来作为一种单独的方言。在接近程度为0.75以上的平面上，北京、济南、西安、扬州、汉口、成都组成了北方方言区；长沙、双峰组成了湘方言区；苏州、温州组成了吴方言区；福州、厦门、潮州组成了闽方言区；梅县、南昌、广州各自独立为一区，梅县为客家方言区，南昌为赣方言区，广州为粤方言区。若取接近程度为0.66为观察平面，这些方言区又组合为三支：一支是北方方言和客家方言；一支是湘方言和赣方言；一支是吴方言、闽方言和粤方言。若取接近程度0.64为观察平面，北方方言与客家方言、湘方言、赣方言就都组合在一起，它们与分布在东南沿海诸省的吴方言、闽方言、粤方言相对峙。显然，这个聚类树形图鲜明地表现了汉语各大方言区的组合情况和它们之间的相互关系。

在各大方言区内部，聚类树形图也表示出了各方言点之间分歧的程度。湘方言内部分歧最大，其次是吴方言内部。北方方言内部则形成三个分支：华北方言和西北方言为一支，江淮方言为一支,西南方言为一支。各次方言内部的接近程度都在0.93以上。闽方言区内部，福州话与厦门话首先组合起来，然后再与潮州话相结合。这些结论与根据非数学方面归纳出来的结果是吻合的。可见，统计数学的方法为汉语方言的研究开辟了新的途径。

我国学者还利用计算机对方言区的人学习普遍话的情况进行了统计分析，摸清了学习普通话的难点，从而促进了推广普通话的工作。云南师范大学利用计算机，通过对四个年级几十个班的学生进行测试，取得了两千多个常用词的统计数据，分析了云南

人在学习普通话时易出现的问题。[①]

他们采用"难度值"来表示某语言成分的易掌握程度，其公式为

$$P=\frac{R}{N}$$

其中，P表示难度值，R是某语言成分的答对人数，N是参加测试的总人数。显然，难度值P越大，答对的人越多，相应的语言成分越容易掌握。

例如，云南人容易读错的声母难度值比较如表1.4.8所示。

表1.4.8　　声母难度值比较

正确音	误读音	难度值%	误读率%	标准差	例　字
n	l	85.56	14.44	6.38	南　难
l	n	83.17	16.83	10.78	来　力
sh	s	77.65	22.35	8.96	师　士
zh	z	73.2	26.8	10.75	至　战
ch	c	71.69	28.31	12.79	成　察
s	sh	71.33	28.67	12.06	岁　肃
c	ch	68.50	31.50	10.45	才　此
z	zh	67.78	32.22	12.86	走　总

由于云南很多方言都能区分n、l，表1.4.8中把n误读为l或把l误读为n 的比例很接近，误读率也不高，只要努力学习，云南人是能够把 n、l 区分开来的。从表1.4.8中还可看出，zh、ch、sh和z、c、s这两组声母的误读情况比较复杂，把z、c、s 误读为zh、ch、sh的误读率明显地高于把zh、ch、sh误读为 z、c、s 的误读率，特别是把z误读为zh，误读率高达32.22%，可见，区分z与 zh 两组声母是云南人学习普通话声母的一个重要问题。

---

① 王渝光，《云南人学普通话的计算机统计研究》，(《云南语言研究》)，1988年，11月。

第5节　计算风格学

风格是人们在交际活动中形成的个人言语特征。这种风格在数量上的表现，就是人们各自的言语在统计特性上的差异。

例如，词长和句长就可以代表人们遣词造句的风格。所谓词长，就是单词中的音节数，所谓句长，就是句子中的单词数。对某个作者的词长和句长进行描述，需要计算平均词长和平均句长。

作者文章中的音节总数被单词总数来除所得的商，就是该作者的平均词长。公式如下：

$$\overline{M}_w = \frac{L_w}{N_w}$$

式中，L_w表示文章中的音节总数，N_w表示单词总数，$\overline{M}_w$表示平均词长。

作者文章中的单词总数被句子总数来除所得的商，叫做平均句长。公式如下：

$$\overline{M}_s = \frac{L_s}{N_s}$$

式中，L_s表示文章中的单词总数，N_s表示句子总数，$\overline{M}_s$表示平均句长。

由于$L_s = N_w$，它们都表示文章中的单词总数，我们可得

$$\overline{M}_s = \frac{L_s}{N_s} = \frac{L_w}{\overline{M}_w N_s},$$

从而有

$$L_w = \overline{M}_w N_s \overline{M}_s$$

有人曾对20位德语作者的22部作品进行过平均词长和平均句长的统计分析。

这22位德语作者是：文学家凯斯特奈(Erich Kästner, 1899—1974)、小说家法拉达(Hans Fallada,1893—1947)，诗人里尔克(Rainer Maria Rilker, 1875—1926)、诗人封丹奈 (Theodor Fentane, 1819—1898)、诗人施托姆 (Theodor Storm, 1817—1888)、小说家和诺贝尔文学奖获得者托马斯·曼(Thomas Mann, 1875—1955)，诗人沙米索 (Adelbert von Chamisso, 1781—1838)，诗人海斯(Hermann Hesse, 1877—1962)、物理学家和诺贝尔物理奖获得者海森堡(Werner Heisenberg,1901—1976)、童话作家豪夫(Wilhelm Hauff, 1802—1827)、物理学家和诺贝尔物理奖获得者爱因斯坦(Albert Einstein,1879—1955)、物理学家索墨菲尔德(Arnold Sommerfeld, 1868—1951)、计算机科学家绍尔(Robert Sauer)、文学家歌德(Johann Wolfgang von Geothe, 1749—1832)、物理学家和诺贝尔物理奖获得者普朗克 (Max Planck, 1858—1947)、文学家霍夫曼(Ernst Hoflmann, 1776—1822)、诗人埃森多夫(Joseph Freiherr von Eichendorff, 1788—1857)、哲学家黑格尔 (Georg Wilhelm Friedrich Hegel, 1770—1831)、无产阶级革命家马克思(Karl Marx,1818—1883)、考古学家施里曼 (Heinrich Schliemann, 1822—1890)。

22部作品中，有文学作品(其中包括小说和诗歌)、文学理论、哲学、经济学、考古学和自然科学 (包括理论物理和计算机科学)。

平均词长和句长如表1.5.1所示。

表1.5.1中，歌德出现了三次，这是因为统计了他的三部作品，风格各不相同：《意大利游记》是散文 (序号14)、《赫尔曼与多罗苔》是叙事长诗(序号15)、《诗与真实》是文学理论著作，其余作者，每人只统计了一部作品，如马克思的《资本论》、黑格尔的《逻辑学》、施里曼的《特洛依考古记》等，为简明起见，表中一般不再列出作品名。

从表1.5.1中可以看出如下的风格特征：

表1.5.1　　　　　　词长与句长

序号	作者名	$\overline{M}_W$	$\overline{M}_S$	序号	作者名	$\overline{M}_W$	$\overline{M}_S$
1	凯斯特奈	1.732	8.432	13	绍尔	2.270	22.600
2	里尔克	1.451	8.747	14	歌德(《意大利游记》)	1.715	22.724
3	法拉达	1.530	10.676				
4	封丹奈	1.724	14.440	15	歌德(《赫尔曼和多罗苔》)	1.575	22.825
5	施托姆	1.631	18.825				
6	托马斯·曼	1.804	18.850	16	普朗克	2.010	23.531
7	沙米索	1.612	19.754	17	霍夫曼	1.721	24.868
8	海斯	1.716	20.011	18	埃森多夫	1.556	24.900
9	海森堡	1.919	20.530	19	歌德(《诗与真实》)	1.686	29.100
10	豪夫	1.645	20.700	20	黑格尔	1.836	31.381
11	爱因斯坦	1.929	21.097	21	马克思	2.021	32.668
12	索墨菲尔德	2.100	21.597	22	施里曼	1.892	42.134

第一、表的上半部大部分是20世纪的作者，如凯斯特奈、法拉达、托马斯·曼、里尔克、海斯等，而表的下半部大部分是18—19世纪的作者，如歌德、埃森多夫、黑格尔、马克思、施里曼等，表1.5.1是按平均句长$\overline{M}_s$递增的顺序排列的，由此可以看出，德语书面语的句子有越来越短的趋势。当然，要精确地论证这个趋势，还要依靠更多的语料进行统计分析。

第二、表1.5.1中，平均句长最高的是人文科学和社会科学家的作品，如施里曼的作品《特洛依考古记》，平均句长为42.134，为小说家凯斯特奈作品平均句长8.432的四倍多。马克思的《资本论》平均句长32.668，而11篇由小说家、文学家写的小说散文(序号1—8，10，17和18）的平均句长是17.292，几乎只有《资本论》平均句长的一半。最有趣的是德国大文豪歌德的作品，他的文学理论著作《诗与真实》的平均句长为29.100，这个数字远远大于他的散文《意大利游记》（平均句长22.724）和叙事长诗《赫尔曼和多罗苔》（平均句长22.825)。由此可以看出，句子长度确实是文

体风格的一个重要标志。

第三、从表1.5.1可知，平均词长与平均句长之间并没有必然的联系。平均词长超过2的共有四位作者，其中普朗克和索墨菲尔德是理论物理学家，绍尔是计算机科学家，马克思是无产阶级革命的理论家。前三位作者用词虽长，但他们的句长只有21—23之间，居于中等。而马克思的作品，词长较长，居第三位；句长也较长，居第二位，看来，马克思是一位善于以长词造长句的作者。霍夫曼、埃森多夫、歌德、黑格尔、施里曼用词平均长度都在 2 以下，句子长度却都在24以上，其中的埃森多夫，词长为1.556，仅大于法拉达（词长为1.530）和里尔克(词长为1.451)，是善于用短词的，而他造的句子平均句长为24.900，居第18位，可见他善于用短词造长句。

通过以上分析，我们可以了解到不同文体、不同作者的风格。

1964年，谢德罗夫（S. Y. Sedelow）提出了"计算风格学"，它是用计算机为手段，对不同作者的风格进行统计、分析、计算、整理的一门新学科①。

计算风格学的产生和发展，使作品的风格的统计研究有了一个科学的理论基础。

计算风格学被成功地应用于"作者考证"的研究中，解决了其中的许多令人棘手的困难问题。

1964年，美国统计学家摩斯泰勒（F. Mosteller）和瓦莱斯(D. L. Wallace)考证出了12篇署名为"联邦主义者"(Federalist)的美国18世纪末期报刊文章的真实作者。作者的侯选人只有两人：一位是美国开国政治家汉弥尔顿（Alexander Hamilton，1757—1804），一位是美国第四任总统麦迪逊(James Madison，1751—1836)。当这两位统计学家开始进行统计分析时，遇到了一个极大

① S. Y. Sedelow, W. A. Sedelow, *A preface to Computational Stylistics*, *System Development Corporation Document*, SP—1354. 1964.

的困难，他们发现，作为风格重要特征的平均句长在这两位作者的已有著作中几乎完全相同，于是，他们只好放弃平均句长这个指标，转而从用词习惯上来找出这两位作者的有区别性的风格特征。他们终于找到了这两位作者在某些虚词的使用上有明显的不同：汉弥尔顿在他的18篇文章中，有14篇用了enough这个词，而麦迪逊在他的14篇文章中，根本不用 enough；汉弥尔顿喜欢用 while，而麦迪逊总是用Whilst；汉弥尔顿喜欢用upon，而麦迪逊则很少用。这样，他们便取得了这两位“候选作者”的风格特征指标。

然后，再把这两位“候选作者”的风格特征指标，与未知的12篇署名“联邦主义者”的文章中相应的风格特征相比较，最后推断这位署名“联邦主义者”的作者就是美国第四任总统麦迪逊。这样，便了结了现代考据学上的这个公案。两位研究者所用的数学方法也得到了学术界的好评。

瑞典学者埃勒加尔（Alvar Ellegärd）利用文章中单词的出现频率来进行“作者考证”，考证的对象是一组写于1769—1772年间的英文信件。这些信件有人认为是一个名叫弗兰西斯（Philip Francis）的英国人写的，此外还有几名“候选作者”。埃勒加尔没有采用如上所述的句法特征而是把这些约有 157 000 个单词的信件从词频的角度与弗兰西斯的著作（共有231 300个单词）和其它候选作者的著作相比较，一共取样100万单词，从中选出458个实词和短语，利用计算机作词频统计，结果发现在弗兰西斯著作中的词频分布情况与这些信件中的词频分布情况最为切合，从而排除了其它候选作者的可能，判定弗兰西斯是这些信件的作者。

埃勒加尔在统计工作中把实词分为两类：一类是对考证目标具有积极意义的实词，称为“积极词汇”，另一类则是“消极词汇”。统计时首先挑选出“积极词汇”，着重从这些积极词汇来研究待考证文章的风格，这种方法被证明是行之有效的。后来被其它学者采用。

由于"作者考证"的情况比较复杂，有时只依据一个方面的风格特征还不足以说明问题，往往需要用到多种因素的指标，从不同的侧面来加以考察。

苏联著名作家肖洛霍夫(M. Шолохов)的名著《静静的顿河》(Тихий Дон)出版后，在1928年就有人说这本书是从另一名不见经传的哥萨克作家克留柯夫(Федор Крюков)那里抄袭来的。到了1974年，一位匿名作者在法国巴黎发表了一本书，断言克留柯夫才是《静静的顿河》的真正作者，肖洛霍夫充其量不过是个合作者罢了，特别是该书的第一、二卷，更是如此。于是，一股怀疑之风又重新刮了起来。①

在这种情况下，捷泽(G. Kjetsaa)等学者决定采用计算风格学的方法来考证《静静的顿河》的真正作者。他们的具体办法是：把《静静的顿河》四卷本同肖洛霍夫和克留科夫两人的其它没有疑问的作品用计算机加以分析比较，以便获得可靠的数据，从而澄清存在的各种疑问。

他们从《静静的顿河》中随机挑选出2 000个句子，再从肖洛霍夫和克留科夫各一篇小说中分别随机挑选出500个句子，一共是三组样本，3 000个句子，输入计算机进行处理，处理步骤如下：

(1) 首先统计句子平均长度，三组样品十分相似。于是再按不同的长度细分成若干组，对三组样本中对应的句子组进行比较，发现肖洛霍夫的小说与《静静的顿河》比较吻合，而克留柯夫的小说与《静静的顿河》则相距甚远。

(2) 第二步，进行词类统计分析。从三个样本中各取出10 000个单词，用χ平方分布的方法，求出词类在三个样本中的分布。统计结果发现，除了代词而外，有六类词肖洛霍夫的小说都与《静静的顿河》相等，而克留柯夫的小说则与之不符。

(3) 第三步，统计各种词类在句子中的不同位置。有人曾经

---

① 钱锋，《计算机和社会科学》，知识出版社，1988年。

研究过，对于俄语这样词序相当自由的语言，词类在句子中的不同位置可以很好地表达文体的风格特点，特别是在句子开头的两个词和句子结尾的三个词往往可以起到区分文体风格的作用。捷泽等人统计了三种样本中句子开头的词类和句子结尾的词类，发现肖洛霍夫的小说与《静静的顿河》十分接近，而克留柯夫的小说则与之有相当距离。

（4）第四步，用计算机作句子结构的分析，统计三种样本中句子的最常用格式，结果发现，肖洛霍夫的小说和《静静的顿河》的最常见句式是用“介词＋体词”开头的句子，而克留柯夫小说的最常见句式，则是用“主语＋动词”开头的句子。

（5）第五步，用计算机统计三种样本中频率最高的15种开始句子的结构，发现肖洛霍夫小说中有14种结构与《静静的顿河》相符合，而克留柯夫小说中只有5种出现于《静静的顿河》中。

（6）第六步，用计算机统计三种样本中频率最高的15种结束句子的结构。发现肖洛霍夫小说中15种结构与《静静的顿河》完全符合，而克留柯夫用以结束句子的结构则与《静静的顿河》完全不同。

根据上述六种不同的统计结果，捷泽等人已有充分的事实来证明，《静静的顿河》确实是肖洛霍夫的作品，不过，由于这是一件世界文学界的大事，他们采取了精益求精的态度，继续进行更大规模的研究，到了1977年，他们已经分析了取自三种不同样本的140 000个单词，其中包括取自《静静的顿河》第四卷的新材料，进一步充实了输入计算机中的语料，这时，捷泽等人才下了一个比较稳健的结论：《静静的顿河》确实是肖洛霍夫的手笔，不过，他在写作时或许参考过克留柯夫的手稿。后来，苏联文学研究者也使用计算机对这个问题进行过考证，得出的结论与捷泽等人的结论相同①。

① 最近，新华社莫斯科1990年5月19日电报道，苏联发现了长篇小说《静静的顿河》的两篇原稿，专家证明，这两篇原稿均出自肖洛霍夫的手笔。这样，《静静的顿河》的作者鉴定这段公案得到完全的澄清。这件事实说明，计算风格学在作者考证方面确实是十分有效的。

我国武汉大学语言自动处理研究组曾用计算机对六个现代作家的12部作品进行过语体风格的研究。12部作品中，有小说四部：老舍的《骆驼祥子》、巴金的《家》、茅盾的《子夜》、赵树理的《三里湾》，有剧本八部：郭沫若的《棠棣之花》、《屈原》、《虎符》、《蔡文姬》、老舍的《茶馆》、《龙须沟》、夏衍的《心防》、《法西斯细菌》。①

如果在某部作品中某个汉字的出现频率高于它在12部作品中的平均频率，那么，把它叫做相对高频字，与相对高频字相反的情况称为相对低频字。作了这样的区分后，从统计结果可以看出，"到、个、几、了、里、令、论、起、且、他、她、一"等字在所有小说中都是相对高频字，而在大部分剧本甚至全部剧本中，它们都是相对低频字。另外一些字，如"吧、哼、啦、呀、吗、你、我"在小说中都是相对低频字，而在大部分剧本中，则是相对高频字。还有一些字，如"的、忽、似"等，在大部分剧本中是相对低频字，而在大部分小说中则是相对高频字，另外一些字，如"好、呢"等，在大部分剧本中是相对高频字，而在大部分小说中，则是相对低频字。这说明了，上述汉字的相对高频和相对低频在小说和剧本之间是呈互补分布的：在小说中具有相对高频，在剧本中就具有相对低频；在剧本中具有相对高频，在小说中就具有相对低频。正是这样一种互补分布，使得小说和剧本在汉字这个平面上区分开来。这样的研究结果对于汉语计算风格学显然是有积极意义的。

第6节　古代语言研究中的统计方法

语言符号的随机性当然也应该存在于古代语言之中，因此，我们就可以用统计方法来研究古代语言。

---

① 彭政策，《汉字与语体的语言风格——汉语计算风格学研究尝试》，(《语言自动处理》)，武汉大学出版社，1988年。

早在1950年，美国语言学家史瓦德士（M. Swadesh）就提出了“语言年代学”(glottochronology)。① 他认为，每一种语言都有一些基本词汇，如人称代词、身体各部分的名称等等，这些基本词汇的变化速度，在很长的时间内大体上是一样的。他选择200个词作为适用于各种语言的基本词汇，经过统计计算出，它们在1000年中保存下来的词汇大约为86%。如果某种古代语言及其发展而成的现代语言的基本词汇有60%是相同或相近的，那么，可根据公式

$$t=\frac{\ln l}{\ln l_0}$$

来计算这种古代语言存在的绝对年代，其中，l_0等于0.86，l是在该现代语言中保留下来的基本词汇的百分比，t是该古代语言存在的绝对年代。

根据条件，$l=0.60$，故得

$$t=\frac{\ln l}{\ln l_0}=\frac{\ln 0.60}{\ln 0.86}\approx 3(\text{千年})$$

也就是说，这种语言从古代算起已经存在3000年了。

如果比较的不是古代语言及其发展而成的现代语言，而是两种由共同原始语分化而来的现代语言，要是这两种现代语言的基本词汇中共同的词的比例为L，那么，这两种现代语言从原始语分化的绝对年代可按公式

$$t=\frac{\ln L}{2\ln l_0}$$

来计算。

例如，比较英语和德语的基本词汇得出：$L=0.82$，由此可知，

$$t=\frac{\ln L}{\ln l_0}=\frac{\ln 0.82}{2\ln 0.86}\approx 1.3(\text{千年})$$

① M. Swadesh, *Salish internal relationship*, «*Int. J. American linguistics*», 16:157—167, 1950.

这意味着，英语和德语是在1300年前即公元6世纪时分化的。

当然，语言的演变的因素是比较复杂的，民族迁徙、民族接触以及其它社会历史因素，经常加快或减慢语言词汇的变化速度，它们对于确定语言发展的绝对年代有着很大的影响，史瓦德士的公式没有考虑到这些复杂因素，当然也就会有一定的局限性。

日本语言学家安本美典和计算机科学家本多正久合作，用计算机把日语词汇与其它有关语言的词汇作统计比较，来研究日语的起源问题。他们还使用了美国语言学家奥斯瓦德(R. L. Oswald)提出的、检验远古语言亲属关系的转移检验法(shift test)，对日语与其它语言偶然的一致进行统计分析，并用英国语言统计学家赫丹 (G. Herdan) 的因子分析法 (factor analysis)，对原始共同语中的因子进行统计分析，提出了"日语诞生波动说"。[①]

"日语诞生波动说" 认为，日语的诞生与印欧语从同一原始分化出来的方式截然不同。日语是以环中国诸语言和古代极东亚洲语为基础，先后通过若干波动，汇合了汉语、柬埔寨语系统语言、印尼语系统语言而形成的，就象若干条涓涓细流汇合成波涛汹涌的大江一样。

当然，这种关于日语起源的理论还要经过进一步的检验，但是，这项研究工作本身说明了，使用统计方法来研究古代语言的演变，不但是必要的，而且也是可能的。

在汉语语音史的研究中，要分析韵部的分合情况。但由于缺乏客观标准，而对同样的韵谱材料，往往见仁见智。韵部同用同到什么程度算是合韵？独用独到什么程度算是分韵？这是长期以来使汉语音韵学家们感到困惑不解的难题。

过去许多音韵学家用枚举例证的方法来研究韵部的分合，常常出现"公说公有理，婆说婆有理"的现象，这是由于语言符号本身的随机性引起的。因为，为了摆脱这种令人棘手的困境，有必

① 安本美典、本多正久，《日本語の誕生》，大修馆书店，1978年。

要使用数理统计的方法。

我国学者在研究“反切”时曾使用过统计方法。如白涤洲在《广韵声纽韵类之统计》①一文中指出，古人做反切，会有两个毛病，一是“同类字太少，随便假借相似的别类字作切”，二是“偶然忽略，误用近似而非同类的字作切”，“我们若不把有这种毛病的字视为例外，严格的依据它考订，态度虽是十分谨严，而实际上反失之呆板”。他说，“依我的浅见，认为用统计方法最适当；把《广韵》一书所用的反切上下字在全书中出现的次数，一一数过，看看哪些字出现的次数多，哪些字出现的次数少，哪几个字简直可以认为是例外，然后再参考前人已用过的方法，斟酌分析，很可以把《广韵》中的声纽韵类，另组成一个系统。”但是，白涤洲实际上只是使用了算术统计的方法，并不足以消除古人做反切时的随机误差。

陆志韦是最早使用概率方法来研究音韵学的中国学者。他在1939年发表的《证广韵五十一声类》②中，提出在统计比较研究时，必须有一个客观标准。他以一个随机相逢概率在样本空间中理论上的实现值作为比较的标准，他把这个标准称之为“几遇数”，用几遇数来跟实际相逢的情况相对照。“凡相逢之数远超乎机率所应得者，因两声类之协合也。凡远不及机率所应得者，因两声类之冲突也。”他还进行了误差估计，确定以±2.5倍作为“远超乎”的标准。这种概率统计方法比算术统计方法更为科学。陆志韦在《广韵说文中间声类转变的大势》、《唐五代韵书跋》、《古音说略》、《古反切是怎样构造的》等文章中，都使用了这种概率统计的方法，在汉语音韵学的研究中，取得了很大成就。

白、陆，二氏的统计方法只是用于“反切”的研究上，如果用于韵语的研究上，则还显得不足。我国学者近年来又提出了进

---

① 白涤洲，《广韵声纽韵类之统计》，（《女师大国学季刊》），1931年，第1期。

② 陆志韦，《证广韵五十一声类》，（《燕京学报》），1939年，第25期。

一步的方法，来研究韵部分合的问题。[①]

这种方法采用统计、计算、判断三个步骤来判断韵部的分合。

统计单位可以分为韵次和字次。以相邻的两韵脚相押一次作为1韵次，韵次记为Y。一个韵脚每押一次，即说它出现1字次，字次记为Z。

例如，一首词中，押L辙（包括A、B、C三韵）的韵脚如下：

$$a_1 b_1 a_2 a_3 b_2 c_1 a_4 b_3$$

a_1，a_2，a_3，a_4表示A韵字，b_1，b_2，b_3表示B韵字，c_1表示C韵字。

a_1和b_1之间押一次韵，即1韵次，b_1和a_2之间也是1韵次，a_2和a_3，a_3和b_2，b_2和c_1，c_1和a_4，a_4和b_3之间也都是1韵次。总共7韵次，其中，aa相押1韵次（即a_2和a_3），ab相押4韵次（即a_1和b_1，b_1和a_2，a_3和b_2，a_4和b_3），ac相押1韵次（即c_1和a_4），bc相押1韵次（即b_2和c_1）。

首尾两韵脚a_1和b_3各算1字次，其余的都算2字次，比如b_1出现2字次，因为它与a_1相押1次，与a_2相押1次。按这样的计算方法，可知A韵字出现7字次（a_1为1字次，a_2，a_3，a_4各为2字次），B韵字出现5字次（b_3为1字次，b_1，b_2各为2字次），c韵字出现2字次（c_1为2字次），总共14字次。

统计L辙内全部字次为：

$$Z_L = Z_a + Z_b + Z_c + Z_Q$$

其中，Z_a，Z_b，Z_c各为A韵字、B韵字，C韵字的字次，Z_Q表示其它辙偶尔押入L辙的字次。

L辙内全部韵次为：

$$Y_L = Y_{aa} + Y_{bb} + Y_{cc} + Y_{ab} + Y_{bc} + Y_{ac} + Y_{LQ} + Y_{QQ}$$

其中，Y_{ab}表示A韵和B韵相押的全部韵次，Y_{bb}，Y_{cc}，Y_{aa}，……的含义按Y_{ab}类推，Y_{LQ}表示其它辙的字偶尔与L辙字相押的韵次，

① 朱晓农，《北宋中原韵辙考》，语文出版社，1989年。

Y_{00}表示其它辙的字偶尔进入L辙并且彼此相押的韵次。

当L辙内的字次和韵次统计结束之后，就可以用概率计算的结果来显示L辙的内部差异情况。这种差异情况，用离合指数来表示。

所谓离合指数就是两韵实际相押比值与理论上相押概率之比。

从理论上说，如果A、B两韵相通，则ab相押的机会只跟A和B两韵字的出现概率有关，于是有

$$P(ab)=\frac{2Z_aZ_b}{(Z_a+Z_b)(Z_a+Z_b-1)}$$

而ab两韵相押的实际比例为

$$R(ab)=\frac{Y_{ab}}{Y_{aa}+Y_{bb}+Y_{ab}}$$

离合指数I可用下式计算：

$$I(ab)=\frac{R(ab)}{P(ab)}\times 100$$

如果A、B完全合成一个韵，则$I\longrightarrow 100$。当离合指数$I\geqslant 100$时，两韵已合并；当$0\leqslant I\leqslant 100$时，$I$值越大，两韵关系越近，$I$值越小，两韵关系越远。

I值的大小可用于预测下一步韵脚的分合情况。一般地说，当$I\geqslant 90$时，可以认为两韵已合并，当$I<50$时，则可认为两韵还未合并，当$50\leqslant I<90$时，两韵似分似合，单靠经验难以判断，可采用"t分布假设检验"的方法来判断。

"t分布假设检验"的具体内容在下面的例子中给出。

下面，以宕辙为例来说明如何使用上述方法。

根据对北宋词人近8000首词的语料的统计结果。可得表1.9.1。

在表1.6.1中，"唐、阳、江、铎"都是宕辙的韵部名称。第一栏数字表示宕辙的总字次和各韵的字次：$Z_{宕}=1362$，$Z_{唐}=361$，

表1.6.1　　韵数统计结果

	1362	唐	阳	江
唐	361	45	102	71T
阳	962	264	339	69T
江	37	7	18	6
铎	2		2	

$Z_{阳}=962$，$Z_{江}=37$，“铎”韵是入声，还出现2个字次，这很可能是个错误，但统计在内并不影响结果。

“楼梯”的左下方的数次是韵次：$Y_{唐唐}=45$（表示“唐”韵和“唐”韵相押45次），$Y_{唐阳}=264$，$Y_{唐江}=7$，$Y_{阳阳}=339$，$Y_{阳江}=18$，$Y_{江江}=6$，$Y_{阳铎}=2$，$Y_{唐铎}=0$，$Y_{江铎}=0$。

这些统计数字之间的关系为

$$Y_{宕}=\frac{1}{2}Z_{宕}=Y_{唐唐}+Y_{唐阳}+Y_{唐江}+Y_{阳阳}$$

$$+Y_{阳江}+Y_{唐铎}+Y_{阳铎}+Y_{江铎}=681$$

$$Z_{唐}=Y_{唐唐}\times 2+Y_{唐阳}+Y_{唐江}+Y_{唐铎}$$

$$=45\times 2+264+7+0=361$$

$Z_{阳}$和$Z_{江}$也跟相应的Y有类似关系。

根据以上的统计数字，可以求出唐江、唐阳、阳江诸韵之间的相互关系。

“楼梯”右上的三个数字是离合指数。这三个离合指数可根据公式求出。

先看唐阳的关系，其相押概率为

$$P_{唐阳}=\frac{2Z_{唐}Z_{阳}}{(Z_{唐}+Z_{阳})(Z_{唐}+Z_{阳}-1)}=\frac{2\times 361\times 962}{1323\times 1322}$$

$$\approx 0.397$$

但实际相押比例为：

$$R_{唐阳}=\frac{Y_{唐阳}}{Y_{唐唐}+Y_{阳阳}+Y_{唐阳}}=\frac{264}{45+264+339}$$

$$\approx 0.4072$$

故离合指数应为

$$I_{唐阳}=\frac{R_{唐阳}}{P_{唐阳}}\times 100=\frac{0.4072}{0.3971}\times 100\approx 102.69$$

按同样的方法，可求出：

$I_{唐江}=71.38$，　$I_{阳江}=69.44$

I值填入"楼梯"的右上方，填入时，不进位舍去小数。

$I_{唐阳}\approx 102>100$，这说明北宋时唐阳二韵已经合并。但$I_{唐江}$和$I_{阳江}$的值太小，很难根据其离合指数说明唐江、阳江相通，这时，需要借助于"t分布假设检验"来判断。

检验阳江是否相通的步骤如下：

(1) 我们需要是单尾检验，因此，零假设和择一假设分别为：

H_0：$\mu=\mu_0$　和H_1：$\mu<\mu_0$

零假设是说阳江已合成一韵，择一假说是说阳江还未合成一韵。

下面来确定零假设正确，还是择一假设正确。

(2) 由统计数字算出标准比值$\mu_0=P_{阳江}=0.0725$

(3) 把宕摄全部统计材料任意分成大致均匀的16组(Z大就多分，Z小就少分)。分组统计如表1.6.2。

(4) 由每组数据算出x_i的值，x_i的计算公式如下：

$$x_i=\frac{Y_{阳江i}}{Y_{阳阳i}+Y_{江江i}+Y_{阳江i}}$$

求得：

$x_1=0$，　$x_2=0.0417$，　$x_3=0.1818$，　$x_4=0.0417$，

$x_5=0.125$，　$x_6=0$，　$x_7=0.1053$，　$x_8=0$，

$x_9=0.0345$　$x_{10}=0$，　$x_{11}=0$，　$x_{12}=0$，

$x_{13}=0.08$，　$x_{14}=0.08$，　$x_{15}=0.0833$，　$x_{16}=0$。

(5) 求出样本均值$\overline{X}$

表1.6.2　　分组统计表

序号	乙	唐唐	唐阳	阳阳	江江	江唐	江阳	其它
1	86	6	25	12	0	0	0	0
2	80	0	16	23	0	0	1	0
3	86	2	17	17	1	2	4	0
4	88	3	17	23	0	0	1	0
5	82	4	11	21	0	2	3	0
6	80	8	19	12	0	1	0	0
7	80	2	19	17	0	0	0	0
8	82	0	21	19	0	1	0	0
9	82	0	12	28	0	0	1	0
10	90	1	17	27	0	0	0	0
11	90	1	16	28	0	0	0	0
12	90	4	13	26	0	0	0	2
13	90	3	17	23	0	0	2	0
14	86	2	16	22	1	0	2	0
15	84	2	15	22	0	1	2	0
16	86	7	13	19	4	0	0	0
总计	1362	45	264	339	6	7	8	2

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i=0.0483$$

（6）求出样本方差S^2

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2=0.0559^2$$

（7）计算统计量t

$$t=\frac{\overline{X}-\mu_0}{\sqrt{\frac{S^2}{n}}}=-1.7327$$

（8）取检验水平$\alpha=0.05$，这意味着江阳已经合成一韵。但我

们错认为它们未合并，犯这种I型错误的可能是5%，$t_{-2\alpha}(n-1) = -1.753$

(9) 最后决定是否取零假设H_0。

若$t < t_{-2\alpha}(n-1)$，则否定零假设H_0；

若$t > t_{-2\alpha}(n-1)$，则取零假设H_0。

现在$t = -1.7327 > t_{-2\alpha}(n-1) = -1.753$，故可取零假设$H_0$，这意味着，我们可以有95%的把握说江阳已经合为一韵。故我们在表1.6.1中在江阳的离合指数69后加T，表示江阳已经相通，合为一韵。

比较t与$t_{-2\alpha}(n-1)$的值可看出，t比$t_{-2\alpha}(n-1)$ 大得并不很多，原因可能是：

① 江韵字仅37个，太少，故引起波动；

② 江阳两韵刚刚合并，或许某些人口头上已不分，某些人还分，或许某些字已不分，某些字还分，也就是说，江阳合韵还处于一个动态的变化当中。

使用同样的"t分布假设检验"方法，可以判断唐江相通，已经合为一韵，故在表1.10.1的唐江离合指数71的后面加T。

由于唐阳相通，江阳相通，唐江相通，最后可得出，北宋时唐阳江三韵已合为一韵的结论。

在汉语语音史研究中，需要处理大量的语言材料，通过这些材料来观察韵辙的分合及其内部差异，从而发现古代汉语中实际的韵部分合情况。在处理这大量的语言材料时，由于语言符号的随机性，数理统计便不是可有可无的东西了。事物本身的随机性决定了非使用统计方法不可。过去在汉语音韵学研究中，对于使用数理统计方法来处理音韵材料的尝试曾经持怀疑的态度，那是由于对语言符号的随机性认识不够造成的。我们相信，将现代数学方法引入到音韵学这门传统的语言学科中，必定会帮助音韵学家们克服传统方法的不足，使音韵学这门古老的学科焕发出青春的活力。

第二章

随机过程与语言符号的冗余性

第1节 语言的使用与马尔可夫链

在第一章中我们曾经指出，语言成分在交际活动中的出现是一个随机事件，但我们在研究这个随机事件时，并没有考虑某一语言成分前后的成分对它的影响。然而在语言的使用时所说出或听到的任何一句话中，这些语言成分之间是前后钩连彼此影响的。如果只把话语中的一个单独的语言成分当作一个随机事件来研究，就难以反映出语言使用的真实面貌。

如果我们把确定语言中字母的出现的试验看成是一个随机试验，把所出现的字母看成是随机试验的结局，那么，语言可以看作是一系列具有不同随机试验结局的链，其中，每一个随机试验的个别结局的概率，依赖于它前面的随机试验的结局。例如，在俄语中，当前面的字母是辅音时，元音出现的概率就增长起来，在字母ч之后，无论如何也不能出现字母ы、я或ю，而主要是出现字母т（在Что这个词中）或и,е等等；如果把空白（记为△）也看成一个字母，那么，在双字音序列△я之后，出现空白△的概率

为0.701，出现字母в的概率为0.157，出现字母з的概率为0.036，出现字母р的概率为0.031，……等等。

因此，为了揭示在交际过程中语言使用的数学面貌，我们必须在连续的语流中，来考察前后钩连彼此相关的各个语言成分之间的概率关系。

俄国数学家马尔可夫(A. A. Марков) 在1913年把普希金叙事长诗《欧根·奥涅金》中的连续字母加以分类，他把元音记为V，把辅音记为C，然后，以3个连续字母为统计单元，统计了这样的三字母序列在《欧根·奥涅金》中的出现次数，得到了如下的元辅音序列表：

$$
\left.\begin{array}{l}
\left.\begin{array}{l}
\left.\begin{array}{l} N(VVV)=115 \\ N(VVC)=989 \end{array}\right\} N(VV)=1044 \\
\left.\begin{array}{l} N(VCV)=4212 \\ N(VCC)=3322 \end{array}\right\} N(VC)=7534
\end{array}\right\} N(V)=8638 \\
\left.\begin{array}{l}
\left.\begin{array}{l} N(CVV)=989 \\ N(CVC)=6545 \end{array}\right\} N(CV)=7534 \\
\left.\begin{array}{l} N(CCV)=3322 \\ N(CCC)=505 \end{array}\right\} N(CC)=3827
\end{array}\right\} N(C)=11362
\end{array}\right\} N=20000
$$

表2.1.1 《欧根·奥涅金》中的元辅音序列表

根据表2.1.1中的数据，可以算出有关元辅音出现的概率。例如，元音出现概率为：

$$P(V)=\frac{N(V)}{N}=\frac{8638}{20000}=0.432$$

元音在辅音后出现的概率为

$$P(VIC)=\frac{N(CV)}{N(C)}=\frac{7534}{11362}=0.663$$

元音在元音后出现的概率为

$$P(V|V)=\frac{N(VV)}{N(V)}=\frac{1044}{8638}=0.128$$

在俄语中，元音在辅音后出现的可能性大于元音在元音后出

现的可能性，马尔可夫的这个表，可以帮助我们找到对这种现象的确切解释。

上面的现象可以概括成随机过程加以研究。

随机过程有两层含义：

第一，它是一个时间的函数，随着时间的改变而改变；

第二，每个时刻上的函数值是不确定的，是随机的，也就是说，每一时刻上的函数值按照一定的概率而分布。

在我们写文章或讲话的时候，每一个字母（或音素）的出现随着时间的改变而改变，是时间的函数，而在每一时刻上出现什么字母（或音素）则有一定的概率性，是随机的，因此，我们可以把语言的使用看成一个随机过程。

这样的随机过程可以用信息论的方法加以研究。

信息论是数学的一个分支，它是研究信息传输和信息处理系统中一般规律的科学。1948年美国数学家申农（C. E. Shannon）从人们长期的通讯实践中，为信息论作了奠基性的工作，三十多年来，这门学科发展极为迅速。

在信息论产生之前，人们对于信息系统的理解是比较肤浅的，一般把携带信息的消息看成是瞬态性的周期性的信号。后来人们把近代统计力学中的重要概念，把马尔可夫随机过程理论以及广义谐波分析等数学方法应用于信息系统的研究中，才看出通讯系统内的信息实质上是一种具有概率性的随机过程，从而得出了一些概括性很高的结论，建立了信息论这个学科。

信息论的研究对象是广义的信息传输和信息处理系统，从最普通的电报、电话、传真、雷达、声纳，一直到各种生物的感知系统，都可以用统一的信息论观点加以描述，都可以概括成这样或那样的随机过程加以研究。

从信息论的角度看来，语言使用这样的随机过程，也就是从语言的发送者通过通讯媒介传输到语言的接收者的过程，如图2.1.1所示。

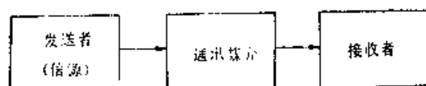


图2.1.1 交际过程示意图

语言的发送者（即信源）随着时间的顺序顺次地发出一个一个的语言符号，语言的接收者也随着时间的顺序顺次地接收到一个一个的语言符号。显而易见，这个过程是时间的函数，而每一时刻的值(即出现什么样的符号)又是随机的。这正是一个确切意义上的随机过程。

在这个随机过程中，所出现的语言成分是随机试验的结局，语言就是一系列具有不同随机试验结局的链。

我们以俄语为例来进行这样的随机试验。在俄语中，如果ъ和ь，е和ё都算为一个字母，词与词之间的空白算为一个字母，那么，俄语字母表就是由32个字母组成的。

如果在随机试验中，各个语言成分的出现彼此独立，不相互影响，那么，这种链就是独立链。

如果在独立链中，每个语言成分的出现概率相等，那么，这种链就叫做等概率独立链。俄语的等概率独立链ϕ_0如下：

оухе рpохьдьш яыхвшхйжтифвнарфсивщтфрпхгпчьнзряс

如果在独立链中，各个语言成分的出现概率不相等，有的出现概率高，有的出现概率低，则这种链就叫做不等概率独立链。

单独俄语字母的概率情况如表2.1.2所示。

考虑到俄语32个字母不等概率，可得到如下的不等概率独立链ϕ_1：

т чыяь оерв одиг збя енвтша ьуемлолйк

在上述独立链信源中，前面的语言成分对后面的语言成分没有影响，是无记忆的，因而它是由一个无记忆信源发出的。

如果在随机试验中，各个语言成分的出现概率不相互独立，每一个随机试验的个别结局依赖于它前面的随机试验的结局，那

表2.1.2　　　　　　　单独俄语字母的概率

字　母	概　率	字　母	概　率	字　母	概　率
空白(△)	0.174	k	0.028	ч	0.012
о	0.090	м	0.026	й	0.010
е, ё	0.072	д	0.025	х	0.009
а	0.062	п	0.023	ж	0.007
и	0.062	у	0.021	ш	0.006
н	0.053	я	0.018	ю	0.006
т	0.053	з	0.016	ц	0.004
с	0.045	б	0.014	щ	0.003
р	0.040	ь, ъ	0.014	э	0.003
в	0.038	г	0.013	ф	0.002
л	0.035				

么，这种链就叫做马尔可夫链。在马尔可夫链中，前面的语言成分对后面的语言成分是有影响的，它是由一个有记忆信源发出的。这正是马尔可夫研究《欧根·奥涅金》的字母序列时所面临的情况。正如马尔可夫所指出的，语言就是由这种有记忆信源发出的马尔

表2.1.3　　　　俄语首字母的概率(出时在△之后时)

字　母	概　率	字　母	概　率	字　母	概　率
п	0.207	я	0.035	е	0.014
н	0.085	з	0.032	э	0.014
и	0.070	т	0.031	л	0.012
с	0.064	ш	0.030	х	0.010
о	0.052	ф	0.029	ц	0.008
в	0.051	р	0.021	ж	0.007
k	0.040	б	0.020	щ	0.003
м	0.038	у	0.020	ю	0.002
д	0.037	г	0.016	я	0.001
а	0.036	ч	0.015		

可夫链。

如果我们只考虑前面一个语言成分对后面一个语言成分出现概率的影响，这样得出的语言成分的链，叫做一重马尔可夫链。

我们把空白(△)也看成一个字母，那么，在△出现时，下一个俄语字母出现概率如表2.1.3所示，这实际上就是俄语首字母的概率分布情况。

俄语字母的一重马尔可夫链ϕ_2如下：

кая всвашный рося ных ковкров

如果我们考虑到前面两个语言成分对后面一个语言成分出现概率的影响，这样得出的语言成分的链，叫做二重马尔可夫链。

我们把空白(△)看成一个字母，那么，在双字母△я之后，俄语字母的概率分布情况如下：

表2.1.4　　△я之后俄语字母的概率

字　母	概　率	字　母	概　率	字　母	概　率
空白(△)	0.701	г	0.004	н	0.001
в	0.157	и	0.003	п	0.001
з	0.036	д	0.002	х	0.001
р	0.031	к	0.002	ш	0.001
щ	0.016	л	0.001		
б	0.009	м	0.001		

俄语字母的二重马尔可夫链ϕ_3如下：

покак постивленный пот дурноскака наконелло зно ствоповил

如果我们考虑到前面三个字母对后一个字母的出现概率的影响，那么，就可得到俄语字母的三重马尔可夫链ϕ_4：

весел враться не сухом и непо и добре

类似地，我们还可以考虑前面四个字母、五个字母……对后面字母出现概率的影响，分别得到四重马尔可夫链ϕ_5、五重马尔

可夫链ϕ_5……等，依此类推。

随着马尔可夫链重数的增大，每一个重数大的俄语字母序列的链，都比重数小的俄语字母序列的链更接近于有意义的俄语文句。

美国语言学家乔姆斯基和心理学家米勒（G. Miller）指出，这样的马尔可夫链的重数并不是无穷地增加的，它的极限就是语法上成立的自然语言句子的集合。这样，我们就有理由把自然语言的句子看成是重数很大的马尔可夫链了。

第2节　语言的熵和语言符号的冗余性

随机过程的一个重要特征是前后符号的相关性，也就是说，从消息的历史，可以预测消息的将来。随着马尔可夫链重数的增大，我们就越能根据前面的语言成分正确地预测下一个语言成分的出现情况，也就是说，随着马尔可夫链重数的增大，我们根据前面的字母预测下一个字母出现的这个随机试验的不肯定性越来越小。至于不是马尔可夫链的句子ϕ_0及ϕ_1，其字母的出现情况是最难预测的，也就是说，每一个字母出现的不肯定性是很大的。

在信息论中，信息量的大小恰恰就是用在接到消息之前，随机试验的不肯定性大小来度量的。随机试验的不肯定性的大小，叫做"熵"（entropy）。

如果我们作某一有n个可能的等概率结局的随机试验（例如，掷骰子，$n=6$），那么，这个随机试验结局的熵就用$\log_2 n$来度量。

这种度量熵的方法是合理的。理由如下：

第一，随机试验的可能结局n越大，这个随机试验的不肯定性程度也就越大，因而它的熵也就越大。

第二，我们做同时包含两个随机试验的复合试验，一个随机

试验有m个可能结局，另一个随机试验有 n 个可能结局（例如，抛硬币时$m=2$，掷骰子时$n=6$），那么，这个复合试验就有$m\cdot n$个可能的等概率结局，也就是说，这个复合试验的熵应该等于$\log_2 m\cdot n$，另一方面，我们又可以认为，这个复合试验的结局的熵应该等于构成这个复合试验的两个随机试验的结局的熵之和，即等于$\log_2 m+\log_2 n$，但根据初等代数知识我们知道：$\log_2 m\cdot n=\log_2 m+\log_2 n$。

可见，复合试验结局的熵，不论是把它看成一个统一的试验还是看成两个随机试验的总和，都是相等的。这个事实证明了我们用$\log_2 n$来度量熵的合理性。

如果随机试验有n个结局，而且，它们是不等概率的，第i个结局的概率为P_i，那么，这个随机试验结局的熵等于

$$-\sum_{i=1}^{n}P_i\log_2 P_i$$

随机试验结局不等概率，减少了这个随机试验的不肯定性(例如，如果骰子的重心有偏斜，那么，掷出来的就常常会是其中的某一点，比如说，“六点”)，因此，有不等式

$$\log_2 n\geqslant-\sum_{i=1}^{n}P_i\log_2 P_i$$

等号当且仅当$P_1=P_2=\cdots=P_n=\frac{1}{n}$时，也就是随机试验的各个结局等概率时才成立。

如果随机试验前面的结局对后面的结局有影响，那么，可得出条件熵：

$$-\sum_{i,j}P[b_i(n-1),j]\log_2 P_{b_i(n-1)}(j)$$

其中，$b_i(n-1)$是由$(n-1)$个结局构成的组合，它后面有第j个结局，$P[b_i(n-1),j]$是这个组合的出现概率，$P_{b_i(n-1)}(j)$是在由前面$(n-1)$个结局构成的组合之后，第j个结局出现的条件概

率。

根据信息论的原理，我们可以分别来计算上述字母链ϕ_0，ϕ_1，ϕ_2，ϕ_3，…中包含在一个字母中的熵。

在字母链ϕ_0中，各个字母等概率不相关，包含在一个字母中的熵由公式$H_0=\log_2 n$来计算，其中n是字母表中的字母数，对于俄语来说，$n=32$，故得：

$H_0=\log_2 32=5$比特/字母

在字母链ϕ_1中，各个字母的出现概率不同，包含在一个字母中的熵由公式$H_1=-\sum_{i=1}^{n}P_i\log_2 P_i$来计算，对于俄语来说，

$$H_1=-\sum_{i=1}^{32}P_i\log_2 P_i=4.35\text{比特/字母}$$

显然，H_1比H_0小。

字母链ϕ_2，ϕ_3，ϕ_4都是马尔可夫链，其中不仅各字母的出现概率不同，而且，每一个字母的出现概率还分别受到直接在它前面的一个、二个、三个字母的影响，这时，包含在ϕ_2，ϕ_3，ϕ_4各字母链的一个字母的熵，分别叫做一阶条件熵、二阶条件熵、三阶条件熵。

一阶条件熵按下面公式来计算：

$$H_2=-\sum_{i,j}P_{ij}\log P_i(j)$$

这里，P_{ij}表示在字母链中一切可能的双字母组合的出现概率，$P_i(j)$表示在前面字母号码为i的条件下，号码为j的字母的出现概率。对于俄语来说，$H_2=3.52$比特/字母，它比H_1小得多。

二阶条件熵按下面公式来计算：

$$H_3=-\sum_{i,j,k}P_{i,j,k}\log_2 P_{ij}(k)$$

这里，P_{ijk}表示一切可能的三字母组合的出现概率，$P_{ij}(k)$表示在号码为i和j的字母之后，号码为k的字母的出现概率。对于俄

语来说，$H_3=3.01$比特/字母，显然，它比H_2小，比H_1更小。

用类似的方法，可以计算包含在字母链的一个字母中的任意阶条件熵。这时，显然，序列H_k是非增的（当各字母等概率时，等号成立）：

$$H_0 \geqslant H_1 \geqslant H_2 \geqslant H_3 \geqslant \cdots \geqslant H_{k-1} \geqslant H_k \geqslant \cdots \rightarrow H_\infty$$

这说明，每在前面追加一个字母，不会使包含在文句的一个字母中的熵有所增加。另一方面，因为包含于字母链的一个字母中的熵，在任何场合都是正的，所以，存在着

$$\lim_{k \to \infty} H_k = H_\infty$$

也就是说，这个序列是有下限的。

通过实验计算表明，在俄语中，从$k=15$开始，H_k在实际上就不再减少而变得稳定起来，这时有

$$H_{15} \approx H_{16} \approx \cdots \approx H_\infty$$

也就是说，如果我们考虑某一字母的前面20个或100个字母，它们对于这个字母的出现概率不会再发生明显的影响。熵H_∞就是包含在字母链的一个字母中的实际信息量，叫做"极限熵"。俄语的极限熵$H_\infty \approx 1$。在俄语中，当字母等概率不相关时，包含于一个字母中的熵$H_0=5$，当字母不等概率，而且前面的字母对后面字母的出现概率有影响时，包含于一个字母中的极限熵H_∞接近于1，可见，字母链中的一个字母的极限熵比所有字母都是等概率不相关时的熵小5倍。即

$$\frac{H_0}{H_\infty} = \frac{5}{1} = 5(\text{倍})$$

熵之所以会减小，是由于语言中各字母之间有相互影响，是由于语言有结构性。在通讯技术中，如果我们给俄语文句编码时，不考虑语言的结构性，使得每一个代码都是等概率不相关的，那么，俄语中就有许多成分显得多余，文句就可以压缩5倍。从这个角度，我们可以说，由于语言的结构性，使得语言中有冗余成分存在。

由于语言的结构性而产生的语言中冗余成分的百分比，叫做冗余度(redundance)，用R表示。冗余度R按下面公式计算：

$$R = 1 - \frac{H_{\infty}}{H_0}$$

例如，对于俄语来说，

$$R = 1 - \frac{H_{\infty}}{H_0} \approx 1 - \frac{1}{5} = 0.8$$

这说明，在任何俄语文章中，大约有80％的字母是由语言本身的结构规定好的，这时，如果我们通过理想的编码，就可以把文章缩减80％（即压缩5倍）。这个事实对于通讯理论和技术都有重要意义。

冗余度可以通过下面简单的试验来估计：我们在任何文章中，读过几个词之后，把文章的后一部分遮住，然后试着去猜测被遮住的第一个字母，然后打开这个字母再猜测下一个字母……如此继续下去，猜中的字母数与被猜测的总字母数之比，就可以作为冗余度的近似值。

语言中的冗余度有四种类型：

1.在书面语中，有不少字母是由语言结构规定好的，根据前面出现的字母，往往可以预测出后面的字母是什么。这就是上文所说的情况。

2.在口语中，情况与书面语类似，当我们漫不经心地用俄语说“Здравствуйте, Александр Алексеевич”时，仿佛是说成“Зрасьсансейч”，这时会俄语的人根据这段口语的上下文仍然可以听懂，知道这句俄语的意思是：“您好！亚历山大·阿列克塞耶维奇”。

3.在文字中，并非构成字母的一切笔画对于辨别这个字母都是必须的。例如，在俄语中，看到-O我们知道是Ю，看到M我们知道是M；如果在一行印刷字母中，我们遮住字母的下面一半，仍能毫不困难地把这一行读懂。

4.在某个语音中，并非语音的一切特征（如音强、音高等）对于辨识这个语音都是必不可少的。例如，发一个[a]音，不论是大人发或小孩发，高音发或低音发，我们都可以辨识出这是[a]。

上述四种类型的冗余度都是必要的和有益的，它保证语言在不理想的条件下（如书面文章中有遗漏，谈话时有嘈杂声，书写的字母不清楚，发音不清晰等），仍能发挥其交际功能。因此，我们不能认为“冗余度”就真的是语言中“冗余”的或不必要的东西。恰恰相反，这种“冗余度”是用语言传递信息时必不可少的。没有冗余度的语言在实际上是无法理解的，因为日常语言总有很大的灵活性，要想理解句子的意思就必须考虑到字母在单词中的位置和单词在句子中的上下文关系。我国著名语言学家李荣先生建议把 redundance 改译为“羡余度”，这是很有道理。事实上，只要语言有结构性就会有冗余性，语言符号的冗余性就是语言的结构性在语言使用过程中的体现。这样看来，语言符号的冗余性也应该是语言符号的一个重要特性，它与语言符号的随机性一样，无时无刻不在语言的使用中表现出来。

在日常语言使用中，语言的冗余度是十分必要的。但是，在通讯技术中，当沿信道传输语言消息时，冗余度往往会造成信道的负荷过重。

研究第一、二种类型的冗余度，有助于解决沿信道传输语言消息的最佳编码问题，从而提高信道的质量。例如，用点和划来给字母编码时，字母的出现频率越高，相应于它的代码序列就应该越是短，这样，就可以减少代码的冗余度。

现代信息论已制定了最佳的编码方法，如费诺（R.M.Fano）编码法。这种编码法是用0和1两个代码来给字母编码的。设有一个消息，其字母表由四个字母组成，这四个字母的概率分别为1/2，1/4，1/8和1/8，我们这样来给这四个字母编码：把这四个字母按概率大小排列起来：1/2，1/4，1/8，1/8，然后，把它们分为两部分，使这两部分的概率之和相等，在第一部分的字母赋

以二进制代码的最初数0，第二部分字母赋以1；接着，再把第二部分分为两部分，使这两部分的概率之和也相等，再分别赋以它们代码数0和1，如此继续下去，直到最后一部分只有一个字母为止。这种手续可由表2.2.1来说明。

表2.2.1　　　　费诺编码法

字母号数	概率	第一次分解	第二次分解	第三次分解	代码
1	1/2	0(概率和为1/2)			0
2	1/4	1(概率和为1/2)	0(概率和为1/4)		10
3	1/8		1(概率和为1/4)	0(概率和为1/8)	110
4	1/8			1(概率和为1/8)	111

使用这种费诺编码法，可以减少冗余度。对于俄语来说，

$$R=1-\frac{H_1}{H_0}\approx 1-\frac{4.35}{5}=0.13$$

可见，这时文句可以压缩13%，如果要对文句作更大的压缩，可按字母组来编码，即不是按单字母编码，而是按双字母、三字母来编码，并用最短的代码序列来记录最常出现的字母组。例如，对于俄语来说，用三字母编码，就可以把文句压缩60%，因为这时有

$$R=1-\frac{H_3}{H_0}\approx 1-\frac{3.01}{5}\approx 0.6$$

研究第三种类型的冗余度（也就是研究字母中哪些笔画对于辨识这个字母是必须的，哪些笔画是冗余的），对于速记文字的设计很有益处。这时，也要利用第一种类型的冗余度，把最常出现的词（如俄语中的что，который,больше等）和字母组合（如俄语中的词尾 -ний，-его，前缀пре-，над-等）用单个符号来表示。

降低第四种类型的冗余度）也就是进行“言语压缩”，可以消除语音的某些冗余特征，使得我们能够更经济地利用信道，提高通讯效率。例如，用言语压缩的方法，可以使电话线路的通过能力增加数百倍。

由此可见，认真地研究语言符号的冗余性，有着十分重大的经济价值。

1951年，申农首先采用信息论的方法测出了英语字母在不等概率独立链中熵H_1，尔后，在实践的迫切要求下，人们又测出了一些印欧语言的熵。到目前为止，英语已测出了9阶条件熵，俄语已测出了14阶条件熵。

测定熵值首先要测出字母的出现概率，对于使用拼音字母的语言来说，只要测出了字母的出现概率，就不难算出相应的熵值H_1。

表2.2.2　　　　英语字母频率表

符　号	p_i	$-\log_2 p_i$	符　号	p_i	$-\log_2 p_i$
空格(△)	0.2	2.32	u	0.0225	5.46
e	0.105	3.25	m	0.021	5.58
t	0.072	3.79	p	0.0175	5.81
o	0.0654	3.93	y	0.012	6.35
a	0.063	3.97	w	0.012	6.35
n	0.059	4.09	g	0.011	6.49
i	0.055	4.18	b	0.0105	6.56
r	0.054	4.20	v	0.008	6.95
s	0.052	4.26	k	0.003	8.35
h	0.047	4.40	x	0.002	9.0
d	0.035	4.84	j	0.001	10.0
l	0.029	5.10	q	0.001	10.0
c	0.023	5.41	z	0.001	10.0
f	0.0225	5.46			

例如，表2.2.2中列出了英语字母的出现概率及其对数，可由此来计算英语的H_1。

由此求得英语的H_1

$$H_1 = -\sum_i P_i \log_2 P_i = 4.03 \text{比特/字母}$$

表2.2.3中列出了德语字母的出现概率及其对数，可由此来计算德语的H_1。

表2.2.3　　　　　　德语字母出现概率表

符　号	p_i	$-\log_2 p_i$	符　号	p_i	$-\log_2 p_i$
空白(△)	0.1442	2.80	o	0.0211	5.57
e	0.1440	2.80	m	0.0172	5.84
n	0.0865	3.53	b	0.0138	6.18
s	0.0646	3.95	w	0.0113	6.45
i	0.0628	3.99	z	0.0092	6.76
r	0.0622	4.00	v	0.0079	6.98
a	0.0594	4.07	f	0.0078	7.00
d	0.0546	4.19	k	0.0071	7.12
t	0.0536	4.22	p	0.0067	7.20
u	0.0422	4.55	j	0.0028	8.48
h	0.0361	4.79	x	0.0008	10.1
l	0.0345	4.85	q	0.0005	11.0
c	0.0255	5.28	y	0.0000…	>14
g	0.0236	5.41			

(ö = oe, ä = ae, ü = ue, β = ss)

由此求得德语的H_1。

$$H_1 = -\sum_i P_i \log_2 P_i = 4.037 \text{比特/字母}$$

现将法语、意大利语、西班牙语、英语、德语、罗马尼亚语、

俄语的不等概率独立链的熵H_1列表比较如下：

表2.2.4　　某些语言的熵H_1

语　　种	符　号　数	熵　H_1	德
法　　语	27个(包括空白)	3.98	拉丁字母
意大利语	22个(包括空白)	4.00	拉丁字母
西班牙语	27个(包括空白)	4.01	拉丁字母
英　　语	27个(包括空白)	4.03	拉丁字母
德　　语	27个(包括空白)	4.037	拉丁字母
罗马尼亚语	27个(包括空白)	4.12	拉丁字母
俄　　语	32个(包括空白)	4.35	斯拉夫字母

汉字是一个相当庞大的字符集，最近开始分册出版的《汉语大字典》，所收汉字超过56 000字，日常生活和报刊杂志上用的汉字大约也有八、九千个，而且这些汉字在书面语中的出现概率又各不相同。因此，要测定在汉语书面语文章中包含在一个汉字中的熵，其计算是十分繁复的。

我国学者用逐渐扩大汉字容量的办法来求汉字的熵值H_1，在实验过程中，我们发现，虽然汉字有将近6万个，但我们在计算汉字的熵值H_1时，并没有必要就这近6万个汉字来计算，只要计算到12 366个汉字就足够了，也就是说，计算汉字熵值H_1的最大汉字容量是12 366个汉字。这是因为：

第一，随着汉字容量的增大，文句中常用汉字的出现概率逐渐趋于稳定，不会再有明显的增大。

例如，常用汉字“的”字的出现概率随着汉字容量增大而变化的情况如下：

表2.2.5　　“的”字的出现概率

汉字容量	1052	1830	4912	5104	5211
出现概率	0.051	0.042	0.041	0.041	0.042

从表2.2.5中可看出，当汉字容量较小时，随着汉字容量由1052扩大到1830，“的”字的出现概率由0.051陡然降到0.042，但随着汉字容量的继续扩大，“的”字的出现概率逐渐稳定于0.042。

汉语中“的”字的出现概率最高，因此，汉字的出现概率$P_i \leqslant 0.042$，即P_i在区间（0，0.042）内取值。在这个区间之内，保持$-P_i\log_2 P_i$随着P_i的增加而增加，如表2.2.6所示。

表2.2.6　　$-p_i\log_2 p_i$随着p_i的增加而增加

p_i	0.001	0.010	0.020	0.030	0.010
$-p_i\log_2 p_i$	0.009966	0.086499	0.112877	0.151767	0.1857504

我们可以作出如下的图像（图2.2.1）。

从图2.2.1及表2.2.6中可以看出，当$P_i \leqslant 0.042$时，汉语中出现概率P_i较高的汉字，它们相应的$-P_i\log_2 P_i$也较高，因而它们对于包含在一个汉字中的熵值H_1的影响也比较大，既然这些常用汉字的出现概率随着汉字容量的扩大而趋于稳定，所以包含在一个汉字中的熵值H_1也将随着汉字容量的扩大而趋于稳定。

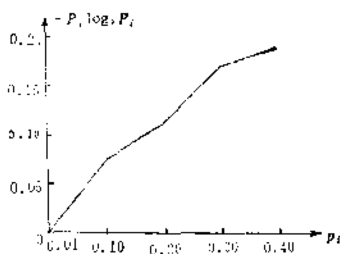


图2.21　$-p_i\log_2 P_i$与P_i的关系

第二，汉语中的非常用汉字的字数虽多，但它们的出现概率极低，随着汉字容量的增大，这些非常用汉字的出现概率还会有所减小，因而包含在一个汉字中的熵值H_1也就会有所减小；而此时随着汉字容量的扩大，文句中又增加了一些新的非常用汉字，从而使包含在一个汉字中的熵值H_1有所增加，这便补偿了由于原来那些非常用汉字的出现概率减小而减小的熵值，使得从总体上看，包含在一个汉字中的熵值H_1变化不大。

那么，究竟当汉语书面语文句中的汉字容量达到多少的时候，

包含在一个汉字中的熵值H_1就不再增加了呢？也就是说，我们能不能求出使包含在一个汉字中的熵值H_1不再增加的最大汉字容量呢？

我们可以借助于齐普夫定律来解决这个问题。①

由齐普夫定律可知，

$$P_r = Cr^{-1}$$

其中，C是一个参数，齐普夫测得$c = 0.1$，r是序号，P_r是序号为r的汉字的频率，这里我仍可以把它看作是汉字的概率。对于$r = 1, 2, \cdots, n$，参数$C = 0.1$，使得

$$\sum_{r=1}^{n} P_r = 1$$

式中的P_r也就是我们这里的P_i，故有

$$\sum_{i=1}^{n} P_i = 1$$

把表示齐普夫定律的公式 $P_r = cr^{-1}$根据我们这里的符号改为

$$P_i = Ci^{-1}$$

代入上式，得

$$\sum_{i=1}^{n} P_i = 1$$

$$\sum_{i=1}^{n} ci^{-1} = 1$$

因$C = 0.1$，从而有

$$0.1 \sum_{i=1}^{n} \frac{1}{i} = 1$$

因此

① 冯志伟，《齐普夫定律的来龙去脉》，（《情报科学》），1983年第2期，第37—42页。

$$\sum_{i=1}^{n}\frac{1}{i}=10$$

$i=1,2,3\cdots,n$，故有

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}=10$$

如果求得了n的值，那么，我们就求得了使文句中各个词出现概率之和为1的最大的汉字容量。欲求n的精确值，可以把调和级数

$$\sum_{i=1}^{n}\frac{1}{i}=1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$$

顺次逐项相加，看加到多少项时其部分和等于10，那么，n的精确值就是多少。但这样做起来运算量太大。这里介绍一种求n的近似值的方法。通过一定的数学推导，我们得到近似公式

$$\sum_{i=1}^{n}\frac{1}{i}\approx\ln n+C$$

式中，ln表示自然对数，$C=0.577215\cdots$，叫做欧拉常数。

由近似公式可得

$$\ln n\approx\sum_{i=1}^{n}\frac{1}{i}-C\approx 10-C=9.422785\cdots$$

由对数换底公式

$$\ln n=\frac{\log n}{\log e}$$

(ln表示自然对数，log表示常用对数，$e\approx2.71828$，是自然对数的底)

得到$\log n=\ln n\cdot\log e=\ln n\times\log 2.71828$

$$\approx 9.422785\times0.43429\approx4.0922213$$

所以，

$$n=12366$$

计算结果告诉我们，当汉字容量大约等于12 366时，这些汉字的出现概率之和为1，如果再增加新的汉字，这些新汉字的出现概率对于整个语言的数学面貌不再会有明显的影响。当根据汉字的出现概率来计算熵值H_1时，如果汉字容量超出12 366，包含在一个汉字中的熵值H_1就不再增加了。

根据这个结论，我们不必在近6万个汉字的容量内来计算汉字熵值H_1，只需在12 366个汉字的容量内来计算就足够了，这样，便大大地简化了汉字熵值的计算，在有限的汉字容量内，我们求得了包含在一个汉字中的熵H_1为9.65比特。这个熵值比起印欧语中包含在一个字母中的熵H_1大得多。

汉字熵值H_1大，说明汉字包含的信息量大，这固然有其优越之处，但是，汉字熵值H_1大，也同时说明其不肯定性程度很大，这将会给通讯技术和中文的计算机处理带来困难，采用现代先进的科学技术来克服这些困难，是我们中国人的光荣使命。目前，我们在这一方面已取得相当大的进展。

我国学者还计算出汉语书面语的冗余度。① 在冗余度的计算公式

$$R = 1 - \frac{H_\infty}{H_0}$$

中，H_∞是极限熵，H_0是最大熵。汉字的最大熵H_0可根据公式$H_0 = \log_{10} n$直接算出。设常用的现代汉字为10 000个，则其中的$H_0 = \log_{10} 10000 = 4$（十进制单位）。但目前对于汉字的条件熵研究得还很差，所以，至今为止，我们还不能直接通过汉字的出现概率及各种条件概率来计算汉字的极限熵H_∞。只有通过间接的办法来估算。

现在，国外已经求出英语字母的极限熵大约在0.28～0.47（十进制单位）之间。我们可以借中英译文为桥梁，根据英语字母的

① 林联合，《关于汉字统计特征的几个问题》，(《语文现代化》)，1980年，第1辑。

极限熵来估算汉语书面文章中汉字的极限熵，从而进一步计算出以汉字为基本单元的汉语书面文章的多余度。

假设同样内容的英语文章与汉语文章的消息量(完全信息)相等，则有

$$H_{\infty}(汉)\times 汉字数 = H_{\infty}(英)\times 英文字母数$$

实验测出，在不计英文的空档时，英语文章中的英文字母数与同样内容的汉语文章中的汉字数之比约为3.7左右，即一个汉字大约相当于3.7个英文字母，而英语的极限熵 H_{∞}(英) 在 0.28到0.47之间，由此推出汉语的极限熵H_{∞}(汉)在1.01到1.74(十进制单位) 之间。汉语书面语的冗余度记为R(汉)，根据冗余度公式

$$R(汉) = 1 - \frac{H_{\infty}(汉)}{H_{0}(汉)}$$

$$= \begin{cases} 1 - \frac{1.04}{4} = 0.74 \\ 1 - \frac{1.74}{4} = 0.56 \end{cases}$$

由此可知以汉字为基本单元的汉语书面语的冗余度约在56%与74%之间，其平均值约为65%。

现在世界上各种语言的冗余度中，计算得比较精确的是英语，柏登 (N. Burton) 和里克里德 (J. Licklider) 两人根据申农的试验方法，通过大量计算求出，英语书面语的冗余度在67%到80%之间，其上下限都略高于汉语书面语的冗余度。①

另外一些实验也表明，印欧语的冗余度略高于汉语，请看表2.2.7

表2.2.7中关于英语冗余度的数据与柏登和里克里德的数据稍有出入。但从这个表可以看出，汉语书面语的冗余度并不算高。

---

① N. Burton, J. Licklider, *Longrange constraints in the statistical Structure of printed English*, *The American Journal of psychology*, 68, No 4, 1955，第650—653页。

表2.2.7　　某些语言的冗余度

文体＼语种	俄语	波兰语	德语	法语	英语	罗马尼亚语
口　语	0.777	0.813	0.792	0.757	0.753	0.801
小　说	0.812	0.791	0.745	0.773	0.818	0.783
科技书刊	0.868	0.866	0.835	0.872	0.875	0.832

我国学者曾采用类似于申农试验的方法，把报纸上的一些句子人为地去掉10％到60％的笔画，要求大学生把省略的笔画填充出来，恢复原来的面貌。如图2.2.2所示：①

他們的眼晴就只看見自已的利益。

滑稽戏是一个专演喜剧的剧种。

他們的讲话一再受到会场的热烈鼓掌。

这简单的称呼体现着生死与共的阶级友爱。

我们种棉花试验有好几年了。②

图2.2.2　冗余度试验样本

试验结果，有一半以上的人在限定的时间内能把笔画省略了55％的文句，完全正确地恢复其原状；在时间宽裕的条件下，少数人能够恢复缺省80％笔画的文句。恢复后的文句为：

他们的眼睛就只看见自己的利益。

滑稽戏是一个专演喜剧的剧种。

他们的讲话一再受到会场的热烈鼓掌。

这简单的称呼体现着生死与共的阶级友爱。

我们种棉花试验有好几年了。

① 曾性初、张履祥，王家柱，《汉语的讯息分析：I.文句中汉字笔画的省略与恢复》，（《心理学报》），1965年4月。

② 这个试验是在1965年进行的，其中用了几个繁体字。不够规范；而且，试验材料的第二行中，漏掉了一个"剧"字。尽管如此，这个试验还是有说服力的。

这次试验反映了书面汉语的冗余度约在55%到80%之间，与根据英语估算的汉语书面语冗余度56%到74%之间悬殊不大。

汉语的冗余度比英语低一些，说明汉语比英语"简练"一些，而"难懂"一些。所谓"简练"一些，就是对同一篇文章，中文将比英文短一些；所谓"难懂"些，指从平均的角度看，文章中对于同样长的字母序列，在语义方面给人们的预示能力差一些，或者说，它的语义更难捉摸一些，语义的不肯定性程度更大一些。

书面文章的冗余度具有两重性。文章的冗余度越高，它就越便于识别和分辨，它的抗错能就越强，因而也就越显得精密，这是冗余度有利的一面。但是，冗余度高，文章的冗余信息就越多，文章就显得不够精练，这是它不利的一面。因此，一种语言文字，它的冗余度不宜过高，也不宜过低，冗余度过高或过低都会给学习和使用带来困难，现存的各种发达语言，都把自己的冗余度在语言的学习和使用的实践中不断地调节到最佳值。

第三章

语言符号的离散性与集合论

第1节　语言符号的离散性

语言符号是由一些离散单元构成的，具有离散性。

我们平时说话时的语流似乎是连续不断的，但在实际上，这些连续不断的语流却是由许多离散单元所组成的。在水平方向上，语流可以被分解为若干个段落，一个段落又可以被分解为若干句子，一个句子又可以被分解为若干短语，一个短语又可被分解为若干单词，一个单词又可分解为若干语素，一个语素又可分解为若干音节，一个音节则是若干个元音辅音音位的组合。在竖直方向上，语流中的各个成分又可引起联想，引出与之属于同一类聚的若干个离散单元来。所以，在连续语流的水平方向和竖直方向上，实际上都是与若干个不同的离散单元联系着的。

例如，“台上坐着主席团”这个句子，在水平方向上，可以分解为

“台／上／坐／着／主席团”

五个离散单元，而在竖直方向上，这个句子中的每一个离散单元

都可以引出一系列与之同类的离散单元。名词"台"引出名词"墙"，动词"坐"引出动词"挂"，名词"主席团"引出名词"月份牌"，便可形成与之平行的句子：

"墙／上／挂／着／月份牌"

在竖直方向上可以联想出的同类的离散单元进一步扩充，就可以形成如下的一系列同类句子：

"台／上／坐／着／主席团"
↓　↓　↓　↓　↓
"墙／上／挂／着／月份牌"
↓　↓　↓　↓　↓
"床／上／躺／着／病人"
↓　↓　↓　↓　↓
"身／上／盖／着／毯子"
↓　↓　↓　↓　↓
"袖口／上／钉／着／钮扣"
↓　↓　↓　↓　↓
"门／上／安／着／电铃"
↓　↓　↓　↓　↓
"山／上／架／着／炮"
↓　↓　↓　↓　↓
"屋／里／摆／着／酒席"

由此我们便能归纳出句型：

$$N_1 + 上(里) + V + 着 + N_2$$

其中，N_1和N_2表示名词，V表示动词。

竖直方向上联想出的离散单元，除了是同类的词之外，还可以是同一个词的不同变化形式。例如，在英语中，由单数名词man 可以联想到其复数形式men，由单数第三人称动词is 可以联想到其复数第三人称形式are，由单数名词 teacher可以联想到其复数形式 teachers。这样，由句子"this man is my teacher"就可以联想出与其平行的句子：

"these men are our teacher"

语言符号的这种离散性，在语流的停延时表现得特别明显，人们往往可以利用语流停延的这种离散性质，来区别语流的不同

含义。①

美国语言学家弗里斯（C.Fries)在他的《英语结构》中提到"5加4乘以6减3"可以有27、17、26、51四种不同的答案，如果根据表达的要求，适当地在语流中配置停延，就可以区别这些不同的答案：

5加4／乘以／6减3＝27

5加／4乘以／6减3＝17

5加4／乘以6／减3＝51

5加／4乘以6／减3＝26

这样的例子还不少，如：

浙江／和/江苏的部分地区有小雨
浙江和江苏的／部分地区有小雨

他说／不下去了
他／说不下去了

他妈的病／还没有好
他妈的／病还没有好

最后一个句子中“他妈的”是骂人话。

利用语言符号性这种离散性可以故意造成阴错阳差，在某种场合反而能起到积极的修辞作用。这里以小学徒报复酒店老板和张士诚的故事来说明这个问题。

老板在酒店门前贴了一副对句，上句是：“酿酒坛坛好做醋缸缸酸”，下句是：“养猪如大山老鼠只只亡”。原意是为了招来顾客，夸他的店里酒、醋有多么好。小学徒施了点小计，在对句上加了两个标点，上句变为“酿酒坛坛好做醋，缸缸酸”，下句变为“养猪大如山老鼠，只只亡”，第二天酒店便没有了生意。

吴晗先生在《朱元璋传》中写张九四作了王爷后要起一个官

① 吴洁敏，停延初探，《语文建设》，1990年，第3期。

名，文人替他起名"士诚"。岂知《孟子》书上有"士，诚小人也"的话，也可以破读成"士诚，小人也"。朱元璋手下的人便以此诋毁文人，说张士诚给人叫了半辈子小人，到死还不明白。

汉语的书面语在词与词之间是连写的，不象印欧语那样留有空白，因此，在汉语书面语中词与词之间的离散特点体现不出来。这种情况，给汉语的自动句法语义分析造成了极大的困难，因此，汉语自动句法语义的第一步便是自动切词，根据词与词之间的离散特征，把相互连在一起的词切开。汉语书面语自动切词的问题我们在第一章第二节中已讲过，兹不赘述。

美国语言学家朱斯(M.Joos)早就指出了语言符号的这种离散性①。他说："数学研究工具一般有两种类型：连续分析（例如，无限小量的计算)或离散分析（例如，有限群理论），而可以称为语言学的那个部门则属于后者，这时，它不容许与连续性有半点儿妥协，因此，语言学可以说成是一个在严格意义上的量子机制，凡是与连续性有关的一切，都得排除于语言学之外。""因此，语言学的范畴是绝对的，是不容许任何妥协的。"他还说，"现在，语言学家把任何语言，也就是任何一个言语行为，看成是由叫做音位的不大数量的基本单位组成的，这些音位在重复出现时被认为是等同的。从物理学的角度来看，hotel这个词对于不同的人或同一个人发音，不可能完全相同地发两次，但从语言学的角度来看，这里却有一个平均数[t]，它始终是同样的，可以不管它们的细微的差别，而把它们看作一个不可分解的语言学原子或范畴，这种原子或范畴，或者是完全等同的，或者是完全不同的。"这里，朱斯十分明确地把语言看成是"不可分解的语言学原子或范畴"离散地结合起来的，因此，他提出用离散数学来研究语言。他说："物理学家利用连续数学来解释言语，如傅利叶分解、自相关函数

---

①朱斯的这些论述，转引自F.Harary H.H.paper, *Toward a general calculus of phonemic distribution*, «*Language*», Vol.33, No.2, 第143—169页

等，而语言学家则与此相反，他们利用离散数学来研究语言。"

朱斯上述关于语言符号的离散性的论述似乎有点儿矫枉过正。语言符号当然具有离散性的一面，因此，我们可以用离散数学来研究它，但是，语言符号也有连续性的一面，特别是在语言的使用中，在语言的交际过程中，我们也可以利用一些连续数学的方法来研究它。朱斯要把"凡是与连续性有关的一切"，"都得排除于语言学之外"，确实是太过分了。事实上，"离散性"和"连续性"都是语言符号本身所具有的性质，不过，在语言的使用中，我们强调语言符号的连续性，用连续数学的方法来研究它，在语言的结构中，我们强调语言符号的离散性，用离散数学来研究它，而语言本身则是离散性和连续性的统一体。当然，朱斯突破了语言学界关于语言符号的"连续性"的传统观念，把"离散性"的观念引入语言学中，从而为采用离散数学来研究语言铺平了道路，他的功绩也是不可抹煞的。

第2节　语言的集合论模型

语言既然是由一些彼此不连续的离散单元组成的，那么，我们就可以把这些离散单元看成集合的元素，采用集合论的方法来研究它，这样，语言研究便与集合论发生了联系。

苏联数学家库拉金娜（О.С.Кулагина）在研究机器翻译的实践中，采用集合论方法来描述语言的某些基本概念，提出了语言的集合论模型。①

首先提出元素x的有限集合$\Xi = x$，元素x称之为"词"。

---

① О, С, Кулагина, Об одном способе определения грамматических понятий на базе теории множеств,«проблемы кцбернстеки»,вып, 1, 1958,第201—214页。

这样，元素x的任何一个有限的有序序列，便可称之为“句子”，记为$A = x_1 x_2 \cdots x_n$。

句子的一切集合分为两个子集：“成立句子”的子集和“非成立句子”的子集。

凡是在形式上正确的句子，都叫做成立句子。所谓形式上正确，是指语法上正确，而不是指语义上正确。因此，在俄语中，“Стол стоит на полу”（桌子立在地板上）和“Тупой куст вразвалку хихикнул”（直译是“迟钝的灌木蹒跚地吃吃笑”，它只是在语法上正确）都是成立句子，而“Он пошел в школа”是非成立句子，因为它在语法上不正确。成立句子的集合，记为$\Theta = \{A\}$。

某一个词的完整的形式系统，也就是，某一个词的词形变化的全部形式的集合，叫做这个词的“域”（Окретность）。例如，对于词Стол(桌子)，有Стол，Стола，Столу，Столом，Столе，Столы，Стлов Стопам，Столами，столах等等，它们构成词стол的一个域。词x的域记为$\Gamma(x)$。

词、成立句子、域这三个概念都是不能在模式中定义的，它们都是从外部提出来的。

库拉金娜从这三个概念出发，演绎地引申出其它概念。

具有给定的域和成立句子的词的集合，称之为“语言”，记为$\Xi(\Gamma, \Theta)$。

彼此不相交的子集的并称为集合Ξ的分划。$\Gamma(x)$把集合Ξ分割为彼此不相交的子集之并，故可得出域的分划，记为Γ分划。

现在引入“等价”的概念。我们说，词x等价于词y，记为$x \sim y$，如果：

1.对于任何一个形如$A_1 x A_2$的成立句子，句子$A_1 y A_2$也成立；

2.对于任何一个形如$B_1 y B_2$的成立句子，句子$B_1 x B_2$也成立。

A_1、A_2、B_1和B_2是任意的句子，它们也可以是不包含任何一

个词的“空句子”。

等价类具有如下的逻辑特点：

1.自反性：$x \sim x$；

2.对称性：如果$x \sim y$，则$y \sim x$；

3.传递性：如果$x \sim y$，且$y \sim z$，则$x \sim z$。

具有上述三个特点的等价，把集合Ξ分割为一系列不相交的子集，这种子集叫做“族”（семейство），两个等价的元素进入同一个族中，而两个不等价的元素则进入不同的族中。词x的族，记为$S(x)$.

例如，我们取俄语句子

(i)Я подшел к окну.

(我走到窗前)

(ii)Прямоугольник,равный окну,очень красиво.

(跟窗子一样大小的那个长方形框子很好看)

在句子(i)中，词окну以两个词串为其环境，一个是“Я подшел к”，一个是空词串。在这个环境中，出现词столу，целовеку等仍得成立句子。在句子(ii)中，词окну以词串“Прямоугольник равный”及词串“очень красиво”为其环境，在这样的环境中，出现词столу，целовеку等仍得成立句子，因此，词окну，столу，человеку等价，属于一个族。

族$S(x)$把集合Ξ分割为彼此不相交子集之和，故可得出族的分划，记为S分划。

这样，我们就得到了用不相交子集系统的形式来表示词的全部集合的两种方法，这就是Γ分划和S分划。在这种场合下，如果我们不管分割出子集的标准是什么，而用彼此不相交子集B_i之并的形式来表示集合Ξ，即

$$\Xi = B_1 \bigcup B_2 \bigcup \cdots \bigcup B_i \cdots \bigcup B_n = \bigcup_{i=1}^{n} B_i$$

那么，我们就把它称之为集合Ξ的B分划。若$x \in B_i$，有时可把B_i

写成$B(x)$。

如果每一个子集只由一个词构成，我们就把这种分划称之为E分划。显然，E分划是B分划的一种特殊情况。

现在我们引入句子A的B结构的概念。取任何一个句子

$$A = x_1 x_2 \cdots x_i \cdots x_n,$$

我们把子集

$$B(x_1)B(x_2)\cdots B(x_i)\cdots B(x_n)$$

的序列，即在给定的B分划中，词x_i所进入的子集的序列，称之为句子A的B结构，记为$B(A)$。

我们取同一个句子

A = раздался звонок(铃响了)

为例，来看看在不同的分划下，这个句子的B结构是怎样的。

1.在E分划下，B结构有形式

$$E(A) = \{\text{раздался}\}\{\text{звонок}\}$$

这种B结构，叫做E结构。

2.在S分划下，B结构有形式

$$S(A) = \left\{\begin{array}{l} \text{раздался} \\ \text{зазвонил} \\ \text{уехал} \\ \text{шел} \\ \text{плакал} \\ \cdots\cdots \end{array}\right\} \left\{\begin{array}{l} \text{звонок} \\ \text{нож} \\ \text{клуб} \\ \text{трамвай} \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \end{array}\right\}$$

这种B结构，叫做S结构。

3.在Г分划下，B结构有形式

$$\Gamma(A) = \left\{\begin{array}{l} \text{раздаться} \\ \text{раздался} \\ \text{раздалось} \\ \text{раздались} \\ \text{раздаются} \\ \cdots\cdots\cdots\cdots \end{array}\right\} \left\{\begin{array}{l} \text{звонку} \\ \text{звонке} \\ \text{звонками} \\ \text{звонки} \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \end{array}\right\}$$

这种B结构，叫做Γ结构。

如果至少有一个成立句子具有某一B结构，那么，这个B结构就是成立的。

我们说，集合B_i与集合B_j是B等价的，记为 $B_i \sim B_j$，如果：

1.对于任何一个形如 $B(A_1)B_iB(A_2)$ 的成立结构，那么结构 $B(A_1)B_jB(A_2)$也成立。

2.对于任何一个形如$B(D_1)B_jB(D_2)$的成立结构，结构$B(D_1)B_iB(D_2)$也成立。

可以看出，前面引入的那种等价的概念是B等价的一种特殊情况，那种等价叫做E等价。

易于检验，B等价也具有自反性、对称性和传递性。

知道了什么是B等价，我们就可以引入“导出分划”的概念了。

设有某个B分划

$$\Xi = \bigcup_{i=1}^{n} B_i,$$

我们这样来构成集合$B'(x)$,使得对于任何的$x \in \Xi$，有

$$B'(x) = \bigcup_{B_i \underset{B}{\sim} B(x)} B_i$$

也就是说，$B'(x)$是在给定B分划下,与$B(x)$处于B等价的一切子集B_i之和。这时，集合$\{B\}$中的任何两个B等价的子集，进入集合$\{B'\}$的同一个子集之中,集合$\{B\}$中的任何两个不等价的子集，进入集合$\{B'\}$的不同的子集之中。显然,在给定B分划下，集合$\{B'\}$构成集合Ξ的一个新的分划：

$$\Xi = \bigcup_{j=1}^{m} B'_j$$

这种由B分划产生的把集合Ξ再分割为一系列不相交子集之并的新的分划，称为B分划的导出分划，记为B'。

用同样的方法，可以从B'产生B'',……等等。库拉金娜证明

了：由导出分划B'产生的导出分划B''与原导出分划B'重合。也就是说，对于任何一个B分划，有

$$B''=B'$$

Γ分划的导出分划，叫做"型"(тип)，记为T。即：$\Gamma'=T$。词x的型，记为$T(x)$.

型与一般语法书中的词类很相近。例如，我们取这样的域：

1.形容词большой，-ая-，-ие等等的一切形式；

2.形容词сильный，ая，-ые等等的一切形式。

显然，这两个域是彼此B等价的，因此，它们属于一个型。词большой的一切形式，词сильный的一切形式，以及俄语其它形容词的一切形式都进入这个型中。可见，型T（большой）很接近于形容词。

这时，如果把不同性的过去时动词也统一于一个型中（第二个型），那么，一切名词也就容易统一于一个型中了（第三个型）。

由此可见，库拉金娜模式中，型与一般语法书中的词类很相近，它可以看作是词类这个概念的数学模型。

下面，我们根据域与族的相互关系来研究库拉金娜模型中一个重要的概念——简单语言的概念。

语言$\Xi(\Gamma,\ \Theta)$称之为简单语言，如果在该语言的域与族之间，满足下面两个要求：

1.对于任何的x，

$$\Gamma(x)\cap S(x)=x$$

成立。换言之，要求进入同一个域中的两个词属于不同的族。

2.如果$x'\in\Gamma(x)$，$y\in S(x)$，那么，

$$S(x')\cap\Gamma(y)\neq\varnothing$$

换言之，如果有词x'进入$\Gamma(x)$中，又有词y进入$S(x)$中，那么，就应该存在某一个词y'，它进入$S(x')$与$\Gamma(x)$之中。

这个要求可以用图来表示。设用实线记$S(x)$和$\Gamma(x)$，用虚线记$S(x')$和$\Gamma(y)$，这时，表示$S(x')$和$\Gamma(y)$的两条虚线就应该相交于某一点y'。

在俄语和其它斯拉夫语言中，第一个要求一般是满足的。例如，对于俄语来说，有

$$\Gamma(\text{стул}) \cap S(\text{стул}) = \text{стул},$$

对于捷克语来说，有

$$\Gamma(\text{stůl}) \cap S(\text{stůl}) = \text{stůl}$$

但俄语不满足第二个要求。

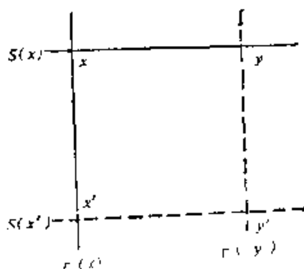


图3.2.1简单语言条件示意图

实际上，стул$\in$Γ(стулья)，реки$\in$S(стулья)，也就是说，词стуля与词стул进入一个域中，又与词ре́ки进入一个族中，即有

$$\Gamma(\text{стул}) \cap S(\text{ре́ки}) = \text{стулья},$$

但是，S(стул)$\cap$Γ(ре́ки) = $\varnothing$，也就是说，没有一个词既能进入词ре́ки的域中，又同时能进入词стул的族中。

这可以用下图表示

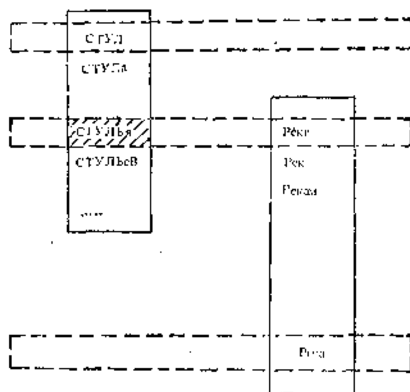


图3.2.2　俄语不满足第2个要求

图3.2.2中，域用实线框表示，族用虚线框表示，族S(ре́ки)与域Γ(стулья)的交用斜线标出。

捷克语也不满足第二个要求。在捷克语中，名词单数有三个族。例如：

阳性{nový žák　　(新同学)
　　 nový stůl　　(新桌子)

阴性nová tužka　　(新铅笔)

中性nové pero　　(新钢笔)

捷克语的名词复数也有三个族，但是，它们是按与单数不同的另一种方式构成的。例如：

人称阳性　　novi žáci　　(一些新同学)

非人称阳性{nové stoly　　(一些新桌子)

阴性　　　 nové tuzky　　(一些新铅笔)

中性　　　 nová pera　　(一些新钢笔)

这时，由于在单数时分别属于两个族的非人称阳性名词和阴性名词在复数时合并为一个族，使得捷克语不能满足第二个要求。事实上，我们有

$$\Gamma(\text{stůl})\ \cap S(\text{tužky})\ = \text{stoly},$$

而

$$S(\text{stůl}) \cap (\text{tužky}) = \varnothing$$

然而，就满足第二个要求的程度来说，捷克语与俄语还是有差别的。列表比较如下：

表3.2.1　　　　捷克语与俄语比较

<table>
<tr><th>语言</th><th>数</th><th>人称阳性</th><th>非人称阳性</th><th>阴性</th><th>中性</th></tr>
<tr><td rowspan="2">捷克语</td><td>单</td><td colspan="2"></td><td></td><td></td></tr>
<tr><td>复</td><td></td><td colspan="2"></td><td></td></tr>
<tr><td rowspan="2">俄语</td><td>单</td><td colspan="2"></td><td></td><td></td></tr>
<tr><td>复</td><td colspan="4"></td></tr>
</table>

在捷克语中，如果我们把一切非人称阳性名词排除出去，那

么，剩下的部分将满足第二个要求，这时，捷克语就可看作简单语言。而俄语在复数时只有一个族，要满足第二个要求是十分困难的。

最后，库拉金娜还提出了"B格式"(B–Конфигурачия)的概念。

联集合Ξ的任意B分划，我们把这样的B结构称为一级B格式，记为$\widetilde{B}_{(1)}$，如果：

1. $\widetilde{B}_{(1)}$含有的元素不少于两个；

2. 存在着B分划的元素B_{a1}，使得B结构$B(A_1)\widetilde{B}_{(1)}B(A_2)$和$B(A_1)B_{a1}B(A_2)$在任何句子$A_1$与$A_2$中，同时成立或者同时不成立。

元素B_{a1}可以在保持结构成立性的条件下替换格式$\widetilde{B}_{(1)}$，我们把它叫做结果元(результурующий элемент)。结果元可能不是唯一的。事实上，如果B_{a1}是格式$\widetilde{B}_{(1)}$的结果元，那么，B分划中与B_{a1}处于B等价的任何元素$B_i(B_i \underset{B}{\sim} B_{a1})$，也可以是格式$\widetilde{B}_{(1)}$的结果元。

用结果元B_{a1}来替换一级B格式，我们便得到一级B结构，记为$B_{(1)}$。

在一般场合下，我们把这样的B结构称为n级B格式，记为$\widetilde{B}_{(n)}$，如果，

1. $\widetilde{B}_{(n)}$含有的元素不少于两个；

2. 存在一个元素B_{an}，使得$(n-1)$级B结构$B(A_1)\widetilde{B}_{(n)}B(A_2)$和B结构$B(A_1)B_{an}B(A_2)$在任何句子$A_1$和$A_2$中，同时成立或同时不成立。其中，不包含n级B格式的B结构$B(A_1)B_{an}B(A_2)$叫做n级B结构。

可见，B格式的定义是递归的：通过$(n-1)$级B结构来定义n级B格式，通过$(n-2)$级B结构来定义$(n-1)$级B格式，……如此等等。

从这样的观点出发，我们来分析下面这个B结构：

B(маленькая)B(девочка)B(долго)B(ласкала)B (кошку)

这是俄语句子

Маленькая девочка долго ласкала Кошку

(小姑娘长时间地抚摩着小猫)

的B结构。

如果我们用B(девочка) 来替换B (маленькая)B(девочка)，得到

B(девочка)B(долго)B(ласкала) B(кошку)，这也是一个成立B结构。但是，这时我们还没有理由认为B (маленькая) B(девочка)这个B结构就是一个B格式，因为我们还没有检查能够进行这种替换的一切环境。

我们再取这样的环境：

B(весьма)B(маленькая)B (девочка)B(стояла)

这是句子

Весьма маленькая девочка стояла.

(很小的女孩站着)

的B结构。

如果我们在这个成立的B结构中，用B (девочка) 来替换B (маленькая) B (девочка)，那么，我们将会得到：

B(весьма) B(девочка) B (стояла)。

这个B结构显然是不成立的。可见，B(маленькая) B(девочка)不是一级B格式。

容易检验，B(весьма) B (маленькая) 是一级B格式，因为B(весьма) B(маленькая) 在一切环境中都可用B(маленькая) 来替换，这时，这个B格式的结果元B_{a1} = B(маленькая)。

如果我们只研究一级B结构，即在其中没有一级B格式的B结构，那么，在任何环境中，B(маленькая)B (девочка) 都可以用B(девочка) 来替换，可见，B(маленькая)B(девочка)是二级B格式，它的结果元B_{a2} = B(девочка)

再继续分析我们的B结构。B(долго) B (ласкала) 是二级B格式，其结果元为B(ласкала)。这样，由原来的那个B结构可得到二级B结构：

B(девочка)B(ласкала)B(кошку)

如果只研究这个二级B结构，那么，在任何环境中，都可用B(стояла)来替换B (ласкала) B (кошку)，也就是用及物动物来替换述宾短语，这样，我们就得到三级B结构：

B(девочка) B(стояла)

从此例可以看出，如果我们把语言成分看成是集合中的一些离散单元，用集合论的方法对其进行分割或分类，就可以为自然语言的自动分析制定出一些严格的语法概念，如族、域、型等等，对自然语言进行形式化的描述。比如上述的格式理论，实际上就是一种归约过程，把复杂的结构一步一步地化为不能再归约的简单结构。这种归约的过程，实际上就是机器翻译中进行句法分析的过程，因此，库拉金娜的集合论模型可以看成是机器翻译句法分析过程的数学模拟。

库拉金娜把她的集合论模型应用到法俄机器翻译系统的研究中，使这个系统能够建立在这种比较完善的数学理论的基础之上，这就为进一步开展机器翻译的研究以及其它的自然语言信息处理的研究，在数学方面提供了一个很好的工具。

罗马尼亚数学家马尔库斯 (S.Marcus) 在库拉金娜工作的基础上，进一步用集合论方法建立了语法性的数学模型，给出了印欧语中阳性、阴性、中性等语法性的严格而清晰的数学描述。由于篇幅的限制，这里就不再多说了。有兴趣的读者可参看马尔库斯的《代数语言学——分析模型》一书①。

① S.Marcus, *Algebraic Linguistics, Analytical Models*, *Academic Press*, 1967。

第四章

语言符号的递归性与公理化方法

第1节 语言符号的递归性

语言符号所构成的句子是无限的，因此，我们不可能枚举出一种语言中所有的句子。在很多场合，对于语言中某一长度有限的句子，往往可以采用一定的办法将其长度加以扩展。例如，下面的句子在英语中都是成立的。

①This is the cat.（这是猫）

②This is the cat that caught the rat.（这是抓老鼠的猫）

③This is the cat that caught the rat that ate the cheese.（这是抓吃乳酪的老鼠的猫）

我们可以在句子①上加上任意个“that从句”，每加一个这样的从句，就构成了一个新的更长的句子。到底能够加多少个 that 从句，只与讲话人的记忆力和耐心有关，而与语言本身的结构无关。我们之所以平时很少说这样的套叠句子，是因为人类心理的短时记忆是有限度的，根据心理学的研究，人们能同时关注到的

事物，短时间内同时记住的东西，以及思维对大脑中同时操纵的元素，都不会超过七个左右①，所以，当一个句子中的成分项目超过七个时，人们就会感到记忆负担过重而不愿说出这样的句子。因此，如果不考虑心理学的因素，仅从语言结构本身来看，我们可以加上无限个that从句而使句子保持成立。

语言符号的这样按同样方式不断扩展的性质，就是语言符号的递归性。

汉语中的宾语从句也可以无限地扩展。例如：

① 我知道小王不知道这件事。

② 我知道小张知道小王不知道这件事。

③ 我知道小李知道小张知道小王不知道这件事。

句子③是合乎语法的，但由于其中的成分项目已超过七个，所以在实际言语中很少会这样说。

上述的英语和汉语例子，都是语言符号的递归性在句法构造方面的表现。

语言在句法构造上所具有的这种递归性，在不同语言里的表现是不尽相同的。在汉语中，句法构造上的递归性突出地表现为句法成分所特有的套叠现象。②

在汉语里，由实词和实词性词语组合成的任何一种类型的句法结构，其组成成分本身可以由该类型的句法结构充任，而无任何形态标志，这就形成了汉语句法成分特有的套叠现象。汉语里由实词和实词性词组组合成的句法结构，有六大类型：主谓结构、偏正结构、述宾结构、述补结构、联合结构、复谓结构。这六种类型的句法结构所特有的成分套叠现象举例如下：

1．主谓结构的套叠

---

① 陆丙甫，《感知对思维的限制》，(《思维科学探索》)，山西人民出版社，1985年。

② 陆俭明，《汉语句法成分特有的套叠现象》，(《中国语文》)，1990年，第2期。

例：我们班上的学生 名字 我 一个 也叫不上来

主语 | 谓语

主语 | 谓语

主语 | 谓语

主语 | 谓语

其中有四个主语套叠。

2.偏正结构的套叠：

例：他 那件 刚买的 还未穿的 新 的确良 短袖 衬衣

定语 | 中心语

定语 | 中心语

定语 | 中心语

定语 | 中心语

定语 | 中心语

定语 | 中心语

定语 | 中心语

其中有七个定语套叠。

例：(他)立刻 满有把握地 用钳子 从墙上 把钉子 慢慢地 一个一个地 拔下来

状语 | 中心语

状语 | 中心语

状语 | 中心语

状语 | 中心语

状语 | 中心语

状语 | 中心语

状语 | 中心语

其中有七个状语套叠。

3.述宾结构的套叠：

例：同意 拟订 一个公约。

述语 | 宾语

述语 | 宾语

其中有两个述语套叠。

4.述补结构的套叠：

例：多得　吃　不完

|述语||　补语　|

|述语||补语|

其中有两个述语套叠。

5.复谓结构的套叠：

复谓结构又可分为连谓结构和递系结构两种。递系结构是一种特殊的复谓结构，其特点是前一项一定是个述宾结构，而其宾语与后一项谓词性词语之间在语义上有直接联系。

连谓结构套叠的例子：

例：下了班　骑车　去姐姐家

|连谓前项||　连谓后项　|

|连谓前项||连谓后项|

其中连谓前项套叠两次。

递系结构套叠的例子：

例：请你　通知老五　来一下

|递系前项||　递系后项　|

|递系前项||递系后项|

其中递系前项套叠两次。

6.联合结构的套叠：

例：（我们准备了）枪枝　和　长矛、　大刀。

|联合前项||　联合后项　|

|联合前项||联合后项|

其中联合前项套叠两次。

上述各种句法成分的套叠现象在汉语中是普遍的、成系统的，是汉语语法的特点之一。

语言的句子是无穷无尽的，而语法规则却是有限的，人们之所以能够借助于有限的语法规则，造出无穷无尽的句子来，其原

因就在于语言符号具有递归性。这些套叠现象正是语言符号的递归性在汉语中的表现。人们在日常生活中使用和理解的句子的范围是无限的，我们之所以能够用有限数目的规则刻画无限数目的句子，正是由语言符号的递归性所使然。

语言机器翻译的实质，就是把源语言中无限数目的句子，通过有限的规则系统，自动地转换为目标语言中无限数目的句子。如果机器翻译规则系统不充分利用语言符号的递归性，要实现这样的转换显然是非常困难的，甚至是不可能的。

现代数学中的公理化方法是研究递归性的有力手段，因此，语言符号的递归性使得语言研究与数学中的公理化方法发生了联系。在这一方面，美国语言学家乔姆斯基的生成语法是对这个问题的最好的说明。

第2节　生成语法的公理化方法

乔姆斯基是当代最有影响的语言学家。1956年，他在研究自然语言的工作中提出了形式语言理论之后，又先后提出了转换语法、生成转换语法的标准理论、生成转换语法的扩充式标准理论、管辖和约束理论等。他的语言学思想就象长江大河一样，不断流动，不断前进，永远也不会停留在一个固定的点上。目前，语言学界对于他的转换语法以及后来提出的关于转换生成语法的各种理论，还有不同的看法，甚至还有不同的争论。但是，他的形式语言理论却成了当代计算机科学的一块重要的基石，已经是人们公认的科学真理了。

乔姆斯基指出，对于自然语言中由于语言符号的递归性而形成的句法结构中的各种一层套一层的套叠现象，可以用有限的规则来加以描述，从而根据有限的规则来生成无限的句子。

例如，对于英语中由“that从句”形成的套叠定语，可以这样来描述：

设X为一个初始符号，S为句子，R为that从句，则有重写规则：

$$X \longrightarrow S, \qquad S \longrightarrow S \frown R$$

这里，“$\longrightarrow$”是重写符号，“$\frown$”是毗连符号。利用这两条规则，便可生成英语中无限个带 that 从句的句子。这有限个刻画语言的规则，叫做“文法”。

这样，乔姆斯基便巧妙地解决了由于语言符号的递归性所产生的语言中无限句子的形式描述问题。

文法是形式语言理论的一个重要概念。所谓文法，就是有限个规则的集合，这些规则能递归地生成数目是潜在地无限的句子。“生成”是文法的核心，它的基础是数学中的公理化方法，阐明了生成语法的公理化方法，我们对于语言符号的递归性，就可以获得更加深入的理解。这里，我们从文法入手，来阐明生成语法的公理化方法。需要说明的是，这里所说的文法，与一般语法书所说的语法不是一码事。它有着严格的形式定义。

形式地说，文法G可定义为一个四元组（V_N，V_T，S，P），即

$$G=(V_N, V_T, S, P)$$

其中

1.V_N是非终极符号的集合，这些符号不能处于生成的终点。

2.V_T是终极符号的集合，这些符号能处于生成的终点，V_N与V_T构成了字母表V，V_N与V_T不相交，没有公共元素，因而有

$$V=V_N \cup V_T \qquad V_N \cap V_T=\varnothing \text{（}\varnothing\text{表示空集合）}$$

V_N中的符号用大写拉丁字母表示，V_T中的符号用小写拉丁字母表示，符号串用希腊字母表示，有时候也可以用拉丁字母表中排在较后面的如w之类的小写拉丁字母表示符号串。

3.S是V_N中的初始符号，它是生成的起点。

4.P是重写规则，其一般形式为

$\varphi \longrightarrow \psi$

这里，φ是V^+中的符号串，ψ是V^*中的符号串，也就是说，$\varphi \neq \varnothing$，即φ中不包含空符号串$\varnothing$。

如果我们用符号“#”来表示符号串的界限，那么，我们可以从初始符号串#S#开始，应用重写规则构成由文法G生成的语言$L(G)$中的成立句子。

利用重写规则$\#S\# \longrightarrow \#\varphi_1\#$（从#S#构成新的符号串$\#\varphi_1\#$），再利用重写规则$\#\varphi_1\# \longrightarrow \#\varphi_2\#$（从$\#\varphi_1\#$构成新的符号串$\#\varphi_2\#$），一直到我们得到不能再重写的符号串为止。这样得到的终极符号串$\#\varphi_n\#$，就是语言$L(G)$的成立句子。

例如，在英语中，有如下的文法

$G = (V_N, V_T, S, P)$

$V_N = (NP, VP, T, N, V)$

$V_T = \{$the, man, boy, ball, saw, hit, took, …$\}$

$S = S$

P:

$S \longrightarrow NP \frown VP$ ①

$NP \longrightarrow T \frown N$ ②

$VP \longrightarrow V \frown NP$ ③

$T \longrightarrow$ the ④

$N \longrightarrow$ man, ball, boy, … ⑤

$V \longrightarrow$ saw, hit, took, … ⑥

利用这些规则，可以从初始符号S开始，生成英语中的成立句子

the man took the ball

the man saw the ball

the man hit the ball

the boy hit the ball

the boy took the ball

……

the man took the boy的生成过程可写成如下形式（后面注明所用规则的号码）：

$$
\begin{array}{ll}
S & \\
NP^\frown VP & ① \\
T^\frown N^\frown VP & ② \\
T^\frown N^\frown V^\frown NP & ③ \\
\text{the } N^\frown V^\frown NP & ④ \\
\text{the man } V^\frown NP & ⑤ \\
\text{the man took } NP & ⑥ \\
\text{the man took } T^\frown N & ② \\
\text{the man took the } N & ④ \\
\text{the man took the ball} & ⑤
\end{array}
$$

这样写出来的生成过程，叫做推导史。

当然，这里的文法只是英语文法的一小部分，生成的语言，也只是英语的一小部分。

又如，我们提出文法

$G=(V_N,\ V_T,\ S,\ P)$

$V_N=\{S\}$

$V_T=\{a,\ b,\ c\}$

$S=S$

P：

$$
\begin{array}{ll}
S \longrightarrow aca & ① \\
S \longrightarrow bcb & ② \\
S \longrightarrow aSa & ③ \\
S \longrightarrow bSb & ④
\end{array}
$$

此文法可生成所谓有中心元素的镜象结构语言，这种语言的句子分为三部分：前面是若干个a及若干个b相毗连，中间是单个的符号c，后面是在c后与前面成镜象关系的若干个a及若干个b的

毗连，即$abcba$，$bbacabb$，$ababacababa$，…用α表示集合$\{a, b\}$上的任意非空符号串，用α^*表示α的镜象，则这种语言可表示为$\{\alpha c\alpha^*\}$。

如果要生成符号串$abbaacaabba$，那么，从S开始的推导史如下：

S	
aSa	③
$abSba$	④
$abbSbba$	④
$abbaSabba$	③
$abbaacaabba$	①

显然，由这个文法生成的语言的符号串的数目是无限的。

下面定义由文法G生成的语言$L(G)$。为此先引入表示V^*上的符号串之间的关系的符号$\underset{G}{\Longrightarrow}$及$\overset{*}{\underset{G}{\Longrightarrow}}$。

如果$\alpha\to\beta$是P中的重写规则，φ_1和φ_2是V^*上的任意符号串，有

$$\varphi_1\alpha\varphi_2\underset{G}{\Longrightarrow}\varphi_1\beta\varphi_2,$$

读为"在文法G中，$\varphi_1\alpha\varphi_2$直接推导出$\varphi_1\beta\varphi_2$"，那么就说，应用重写规则$\alpha\to\beta$于符号串$\varphi_1\alpha\varphi_2$，得到了符号串$\varphi_1\beta\varphi_2$。可见，当应用某个单独的重写规则从第一个符号串得到第二个符号串的时候，$\underset{G}{\Longrightarrow}$表示这两个符号串之间的直接推导关系。

假定α_1，α_2，α_3，…，α_{m-1}，α_m是V^*上的符号串，并且，$\alpha_1\underset{G}{\Longrightarrow}\alpha_2$，$\alpha_2\underset{G}{\Longrightarrow}\alpha_3$，…，$\alpha_{m-1}\underset{G}{\Longrightarrow}\alpha_m$，那么，就写为$\alpha_1\overset{*}{\underset{G}{\Longrightarrow}}\alpha_m$，读为"在文法$G$中，$\alpha_1$推导出$\alpha_m$"。简言之，如果应用$P$中的若干个重写规则由$\alpha$得到$\beta$，那么就说，对于这两个符号串，有

$$\alpha\overset{*}{\underset{G}{\Longrightarrow}}\beta$$

$\underset{G}{\Longrightarrow}$表示$\alpha$与$\beta$这两个符号串之间的推导关系。

这样，由文法G生成的语言$L(G)$可定义如下：

$L(G)=\{w|w在V_T^*中，并且S\overset{*}{\underset{G}{\Longrightarrow}}w\}$。

其意思是：对于一切的符号串w的集合，w在V_T^*中，并且$S\overset{*}{\underset{G}{\Longrightarrow}}w$，那么，符号串w的集合就是由文法$G$生成的语言$L(G)$。由此可见，一个符号串处于$L(G)$中，要满足两个条件：

1.该符号串只含有终极符号；

2.该符号串能从初始符号S推导出来。

同一语言可以由不同的文法来生成。如果$L(G_1)=L(G_2)$，则文法G_1等价于文法G_2。

前面所定义的文法$G=(V_N, V_T, S, P)$，其重写规则为$\varphi\to\psi$，并且要求$\psi\neq\phi$。这样定义的文法，其生成能力太强了。为此，乔姆斯基给文法加上了程度各不相同的一些限制，从而得到了生成能力各不相同的四类文法。

限制1：如果$\varphi\to\psi$，那么，存在A，φ_1，φ_2，ω，使得$\varphi=\varphi_1A\varphi_2$，$\psi=\varphi_1\omega\varphi_2$。

限制2：如果$\varphi\to\psi$，那么，存在A，φ_1，φ_2，ω，使得$\varphi=\varphi_1A\varphi_2$，$\psi=\varphi_1\omega\varphi_2$，并且$A\to\omega$。

限制3：如果$\varphi\to\psi$，那么，存在A，φ_1，φ_2，ω，a，Q，使得$\varphi=\varphi_1A\varphi_2$，$\psi=\varphi_1\omega\varphi_2$，$A\to\omega$，并且$\omega=aQ$或$\omega=a$，因而，$A\to aQ$或$A\to a$。

限制1要求文法的重写规则全部具有形式$\varphi_1A\varphi_2\to\varphi_1\omega\varphi_2$，这样的重写规则在上下文$\varphi_1\to\varphi_2$中给出$A\to\omega$。显然，在这种情况下，$\psi$这个符号串的长度(即$\psi$中的符号数)至少等于或者大于$\varphi$这个符号串的长度(即$\varphi$中的符号数)，如果用$|\psi|$和$|\varphi|$分别表示符号串$\psi$和$\varphi$的长度，则有$|\psi|\geqslant|\varphi|$。由于在重写规则$\varphi_1A\varphi_2\to\varphi_1\omega\varphi_2$中，每当$A$出现于上下文$\varphi_1—\varphi_2$中的时候，可以用$\omega$来替换$A$，

因此，把加上了限制1的文法叫做上下文有关文法（context--sensitive grammav）或1型文法（type 1 grammar）。

限制2要求文法的重写规则全都具有形式$A \to \omega$，这时上下文$\varphi_1 — \varphi_2$是空的，在运用重写规则时不依赖于单个的非终极符号A所出现的上下文环境。因此，把加上了限制2的文法叫做上下文无关文法（context-free grammar）或2型文法（type 2 grammar）。

限制3要求文法的重写规则全都具有形式$A \to aQ$或$A \to a$，其中，A和Q是非终极符号，a是终极符号。这样的文法叫做有限状态文法（finite state grammar）或3型文法（type 3 grammar），有时也叫做正则文法（regular grammar）。

没有上述限制的文法，叫做0型文法（type 0 grammar）。

显而易见，每一个有限状态文法都是上下文无关的；每一个上下文无关文法都是上下文有关的；每一个上下文有关文法都是0型的。乔姆斯基把由0型文法生成的语言叫做0型语言（type 0 language），把由上下文有关文法。上下文无关文法和有限状态文法生成的语言分别叫做上下文有关语言（context-sensitive language）、上下文无关语言（context-free language）和有限状态语言（finite state language），也可以分别叫做1型语言（type 1 language）、2型语言（type 2 language）和3型语言（type 3 language）。

由于从限制1到限制3的限制条件是逐渐增加的，因此，不论对于文法或对于语言来说，都存在着如下的包含关系：

$$0型 \supseteq 1型 \supseteq 2型 \supseteq 3型$$

可图示为图4.2.1。

例如，有文法$G = (V_N, V_T, S, P)$

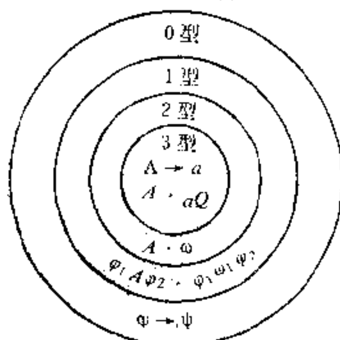


图4.2.1 文法和语言的分类

$V_N=\{S, A, B, C\}$

$V_T=\{a, b, c\}$

$S=\{S\}$

P:

$S\to ABC$ ①

$A\to aA$ ②

$A\to a$ ③

$B\to Bb$ ④

$B\to b$ ⑤

$BC\to Bcc$ ⑥

$ab\to ba$ ⑦

这个文法可生成终极符号串$b^n a^m cc(n\geqslant 1, m\geqslant 1)$。

不难看出，规则⑦$ab\to ba$是0型规则，因此，这个文法是0型文法。如果去掉规则⑦，那么，就得到一个1型文法，因为规则⑥$BC\to Bcc$是1型规则。如果去掉规则⑦和⑥，就得到一个2型文法，因为规则①$S\to ABC$和规则④$B\to Bb$是2型规则。如果去掉规则⑦、⑥、①、④，就得到一个3型文法，因为剩下的规则②$A\to aA$，规则③$A\to a$和规则⑤$B\to b$都是3型规则。

可见，任何的3型文法，一定包含在2型、1型、0型文法中，任何的2型文法，一定包含在1型0型文法中，任何的1型文法，一定包含在0型文法中。

上述四种类型的文法及其所生成的语言的卓越见解，是乔姆斯基对于形式语言理论的最为重要的贡献，在计算机科学中，人们把它称之为乔姆斯基分类(Chomsky classification)。

乔姆斯基的形式语言理论，对于计算机科学有重大意义。乔姆斯基把他的四种类型的文法分别与图灵机、线性有界自动机、后进先出自动机及有限自动机等四种类型的自动机联系起来，并证明了文法的生成能力和语言自动机的识别能力的等价性的四个重要结果，即：

1.若一语言L能为图灵机识别，则它就能由0型文法生成，反之亦然；

2.若一语言L能为线性有界自动机识别，则它就能由1型(上下文有关)文法生成，反之亦然；

3.若一语言L能为后进先出自动机识别，则它就能由2型(上下文无关)文法生成，反之亦然；

4.若一语言L能为有限自动机识别，则它就能由3型(有限状态)文法生成，反之亦然。

乔姆斯基的上述结论，提供了关于语言生成过程与语言识别过程的极为精辟的见解，这对于计算机的程序语言设计、算法分析、编译技术、图象识别、人工智能等领域的研究，都是很有用处的，因而在计算机界产生了很大的影响。特别是在计算机科学家们发现，算法语言ALGOL60中使用的巴库斯-瑙尔范式 BNF 恰好与乔姆斯基的上下文无关文法CFG等价之后，不少学者都投入了上下文无关文法的研究，精益求精，成绩斐然。

在语言学界，常常把上下文无关文法叫做短语结构文法，不少学者都注意研究短语结构文法的生成能力，并提出了有效的方法来改进它，使之更适合于自然语言的描述，而且，在许多机器翻译系统中都采用短语结构文法作为描写自然语言的基本方法。

上面，我们说明了文法的基本概念，并且把文法定义为$G=(V_N, V_T, S, P)$四元组。这一定义是生成语法的关键，那么，文法这一定义在数学上的根据是什么呢？

我们认为，文法这一定义的根据是数学上的公理系统理论。

象初等几何学这样的公理系统是怎样建立起来的呢？

首先，它要有一系列的公理，公理是不需要说明的大家公认的显而易见的真理，以这些公理作为建立几何学系统的出发点；其次，它要有一系列的推导规则，以便从公理出发，一步一步地推导出各种定理来。

例如，我们可以这样来证明三角形ABC各内角之和等于180°，

如图4.2.2所示。

首先，根据“从直线外一点仅能作一条直线与该直线平行”这一公理，通过C作一直线CE平行于AB，再根据同位角相等的规则，可知$\angle ECD = ABC$，又根据内错角相等的规则，可知$\angle ACE = \angle BAC$，再根据等量变换的规则，由

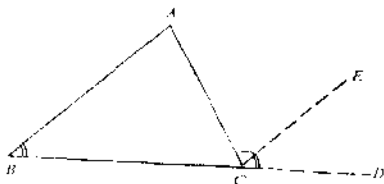


图4.2.2 三角形各内角之和等于180°

$$\angle ECD + \angle ACE + \angle ACB = 180°,$$

推出

$$\angle ABC + \angle BAC + \angle ACB = 180°,$$

从而证明了三角形ABC各内角之和为180°。

这个证明的逻辑结构如下：

公理：　　从直线外一点只能作一条直线与该直线平行。

↓

推理规则：①内错角相等；②同位角相等；③等量替换。

↓

定理：　　三角形各内角之和等于180°。

可以看出，证明的逻辑结构是：从公理出发，运用若干条推理规则，最后得出定理。

我们是不是也可以采用类似的方法来刻画语言，从而从形式上来描述语言的生成过程呢？

我们知道，从本质上说，语言是一个无限集。如果我们取语言这个集合的某个真子集合，那么，这个真子集合总是与语言这个集合等价。例如，在讲某一语言的某一社会集团中，我们取该社会集团某一成员所讲的话为该语言的真子集合，显然，这个成员所讲的话总是与这种语言等价，也就是说，在语言这个集合中，不同大小的集合之间可以建立一一对应关系，“部分小于全体”

这个在有限集内适用的规律，在语言这个集合中，却变成了“部分等于全体”，而这正是无限集的特征。因此，我们说，语言在本质上是一个无限集。

乔姆斯基对这个问题的看法还要深刻得多，他指出，早在19世纪初，德国杰出的语言学家和人文学者洪堡德（W. V. Humboldt）就观察到“语言是有限手段的无限运用”，但是，由于当时尚未找到能揭示这种理解所含的本质内容的技术工具和方法，洪堡德的论断还是不成熟的。那么，究竟应该如何来理解语言是有限手段的无限运用呢？乔姆斯基指出：“一个人的语言知识是以某种方式体现在人脑这个有限的机体之中的，因此语言知识就是一个由某种规则和原则构成的有限系统。但是一个会说话的人却能讲出并理解他从来未听到过的句子及和我们所听到的不十分相似的句子。而且，这种能力是无限的。如果不受时间和注意力的限制，那么由一个人所获得的知识系统规定了特定形式、结构和意义的句子的数目也将是无限的。不难看到这种能力在正常的人类生活中得到自由的运用。我们在日常生活中所使用和理解的句子范围是极大的，无论就其实际情况而言还是为了理论上描写的需要，我们完全有理由认为人们使用和理解的句子范围都是无限的。”①

那么，怎样来刻画语言这个无限集的成分组成情况呢？

我们可以把语言中所有的元列成一个表，进行简单枚举。例如，

$$L=\{\phi,\ a,\ b,\ aa,\ ab,\ \cdots\}$$

这样的刻画办法，把后面一大部分东西省略掉了，后面未列出的部分，只好由我们根据表中给出的少量的元去想象，这样的刻画办法显然是不好的。它不能体现“有限手段的无限运用”这一原则。

---

① N. Chomsky，乔姆斯基序，《乔姆斯基理论介绍》，中文本，黑龙江大学出版社，1982年。

由于语言符号具有递归性，而在数学中，能够体现“有限手段无限运用”的最好办法是递归，所以，我们可采用递归来刻画语言这个无限集。

我们来研究在$\{a, b\}$上的一切镜象符号串的集合M。一个镜象符号串可分为左半和右半两部分，右半包含的符号序列与左半包含的符号序列相同而顺序相反。例如，$aaaa$，$abba$，$babbab$，$bbabbabb$都是镜象符号串，但$babb$，$aaab$不是镜象符号串。显然，这种镜象符号串的集合M是一个无限集，我们用下面的递归定义来刻画它：

(1)
- (i) $aa \in M \wedge bb \in M$；
- (ii) $(\forall x)(x \in M \rightarrow (axa \in M \wedge bxb \in M))$；
- (iii) 除具有性质(i)、(ii)的元之外，M不包含其它的元。

(i)叫做递归定义的基底，它说明，对于特定的符号串aa和bb，$x \in M$为真。

(ii)叫做递归步骤，它说明，对于任意的符号串x，如果$x \in M$为真，那么，在x两侧毗连a或毗连b所构成的符号串也为真。

(iii)叫做限制，它排除了不满足(i)和(ii)的$x \in M$的其它一切情况。如果没有这个限制，递归定义就可以描述满足条件(i)和(ii)，但同时还可能包含其它元素的集合。

应该注意的是，在递归定义中，基底是必不可少的，有了基底，我们就可以取递归步骤的适当的替换值，采用数理逻辑中的取式推理法(modns ponens)，推出这个替换值的结果，从而证明某个符号串是否在M中。

数理逻辑中的取式推理法其推论形式如下：

$$\begin{array}{l} p \rightarrow q \\ \underline{\quad p \quad} \\ \therefore\ q \end{array}$$

这种推理的形式无疑是正确的。

现在，我们从(1)中的基底和递归步骤，来证明$bbaabb \in M$。

1. $aa \in M \wedge bb \in M$ 基底

2. $(\forall x)(x \in M \rightarrow (axa \in M \wedge bxb \in M))$ 递归步骤

3. $aa \in M$ 对1简化

4. $aa \in M \rightarrow (aaaa \in M \wedge baab \in M)$ 对2全称赋值

5. $aaaa \in M \wedge baab \in M$ 对3、4作取式推理

6. $baab \in M$ 对5简化

7. $baab \in M \rightarrow (abaaba \in M \wedge bbaabb \in M)$

取$baab$为替换值对2全称赋值

8. $abaaba \in M \wedge bbaabb \in M$ 对6、7作取式推理

9. $bbaabb \in M$ 对8简化

如果没有基底，从递归步骤只能推出一系列的条件语句，而得不到证明。

采用这种递归的办法，任何具有镜象结构的符号串，都可以通过基底和递归步骤，采用适当的逻辑推理来证明它必属于集合M。在这种情况下，我们可以把基底看成是原先给定的有限数目的命题，把递归步骤看成是一套特定的有限数目的规则，也就是说，采用递归，我们可以从原先给定的有限数目的命题出发，反复运用一套特定的有限数目的规则，推导出无限数目的外加命题。这就是"有限手段的无限运用"。可见，递归确实是刻画无限集的好办法。

为了应用递归来刻画语言这个无限集，我们提出公理系统的定义。

一个公理系统是一个有序三元组(A, S, P)，其中，

1. A是符号的有限集，叫做字母表；

2. S是A上的符号串的集合，叫做公理；

3. P是在A^*的符号串上的n位关系的集合，$n \geqslant 2$(即P中的n元组至少必须是有序对)，P的元叫做生成式或推理规则。

由此，我们易于看出，递归定义很象一个公理系统，其中，基底相当于公理，递归步骤相当于推理规则，递归定义所刻画的

集合的元，除了那些由基底给出的元之外，就构成了这个公理系统的定理。

现在我们说明，如何使用生成式来推出定理。

给定一个公理系统(A, S, P)，如果$(x_1, x_2, \cdots, x_{n-1}, x_n)$是$P$中的一个生成式，那么，我们就说，$x_n$是由$(x_1, x_2 \cdots, x_{n-1})$导出的，我们可把$(x_1, x_2, \cdots, x_{n-1}, x_n)$等价地记为$x_1, x_2, \cdots, x_{n-1} \to x_n$。

给定一个公理系统(A, S, P)符号串$y_1, y_2, \cdots, y_m$的线性有序序列叫做y_m的一个推导或证明，当且仅当这个序列中的每一个符号串或者是公理，或者是用P中的一个生成式从该序列中它前面的一个或多个符号串导出的符号串。如果在给定的公理系统中，存在着y的一个推导，那么，y就叫做该公理系统的一个定理。

例如，(1)中在$\{a,b\}$上的镜象符号串的递归定义可这样地解释为一个公理系统(A, S, P)：

$$(2) \quad \begin{cases} A = \{a, b\} \\ S = \{aa, bb\} \\ P = \{(x, y) \in A^* \times A^* \mid y = axa \vee y = bxb\} \end{cases}$$

这时，生成式是如下的有序对集合：

$\{(\phi, aa), (\phi, bb), (a, aaa), (a, bab), (b, bbb), (aa, aaaa), \cdots\}$

换一种记法，这个集合可写为：

$\{\phi \to aa, \phi \to bb, a \to aaa, a \to bab, b \to bbb, aa \to aaaa, \cdots\}$

在公理系统(2)中，我们可看出，线性有序序列

$bb, abba, aabbaa$

是$aabbaa$的一个推导，因为这个序列的最后一个符号串是从它直接前面的符号串用生成式$abba \to aabbaa$推出的，而$abba$是从bb用生成式$bb \to abba$推出的，由于bb是公理，所以，$aabbaa$就是公理系统(2)中的一个定理。

序列

bb，$baab$

不是一个推导，因为$baab$不是从bb用P中的规则推出的。但这并不意味着$baab$不是一个定理，因为在公理系统(2)中，可以存在一个推导，使得$baab$是该推导的最后一行。这个推导是

aa，$baab$

因此，$baab$也是一个定理。

由此可知，推导的第一行必须是公理，因为在第一行前，没有什么能够推出它。因此，下面的序列

ab，$aaba$，$baabab$

不是推导，因为ab不是公理。

一个推导可以只包括一行，这一行必是公理。

(2)中生成式P的集合是所有形式为(x, axa)和(x, bxb)的一切有序对的无限集。其中，x是变量，它的值是A^*中所有的符号串，因此，P可包含如(a, aaa)和$(ab, babb)$这样的生成式，但这样的生成式，在(2)的公理系统中，当从给定的一套公理出发来推导任何的定理时，实际上是从来不使用的。另外，由于x是一个变量符号，而不是字母表A中的一个元，所以，(x, axa)和(x, bxb)本身并不是生成式，而是生成式格式或构造生成式的公式。生成式格式这个有限集，刻画了生成式的无限集，其中，变量x可用A^*中的任何符号串来代表。

前面关于"三角形内角之和等于180°"的那个证明的逻辑结构是：从公理出发，运用若干个推理规则，最后得出定理。这样的逻辑结构，正是公理系统理论在建立初等几何学的公理系统中的体现。

我们还可以把公理系统的定义扩展到容许字母表中出现两类字母表：一类叫基本字母表，一类叫辅助字母表，它们是不相交的两个集合。两类字母表中的符号都可以出现在推导的行中，但是定理中只包含来自基本字母表中的符号。这一种有两类不相交字母表的公理系统，叫做扩展公理系统。定义如下：

一个扩展公理系统是一个有序四元组(A, B, S, P)，其中：

1. A是辅助字母表符号的有限集；

2. B是基本字母表符号的有限集，A与B不相交；

3. S是 $(A\cup B)^*$ 上的符号串的集合，即公理，S可由公理的有限集来刻画；

4. P是在 $(A\cup B)^*$的符号串上的n 位关系的集合 $(n\geqslant 2)$，叫做生成式或推理规则。P可以用生成式格式的有限集来刻画，如果 $(x_1, x_2, \cdots, x_{n-1}, x_n)$ 是 P 中的生成式，我们就说，x_n 是从 $x_1, x_2, \cdots, x_{n-1}$ 中导出的，可等价地记为 $x_1, x_2, \cdots, x_{n-1}\rightarrow x_n$。

在扩展公理系统中，我们有必要把推导和证明区别开来，因为并不是每一个推导都是以定理为结尾的。

给定一个扩展公理系统 (A, B, S, P)，符号串的线性有序序列$y_1, y_2, \cdots, y_m$叫做y_m的一个推导，如果每一个符号串是一个公理或者每一个符号串是用 P 中的一个生成式从该序列中这个符号串前面的一个或多个符号串导出的。

给定一个扩展公理系统 (A, B, S, P)，符号串叫做一个公理，如果：

1. 它是 (A, B, S, P) 中y的一个推导，并且，

2. $y\in B^*$

当y是一个定理时，y的推导就叫做y的证明。

我们可以看出，每一个公理系统也是一个以零集为辅助字母表的扩展公理系统，但并非每一个扩展公理系统都是公理系统，带有非零集辅助字母表的扩展公理系统是一个真扩展公理系统。

例如，定理为$\{a,b\}$上的镜象符号串的真扩展公理系统如下：

$$
(3)\quad \begin{cases}
A=\{m\} \\
B=\{a,\ b\} \\
S=\{m\} \\
P: \\
\qquad \alpha M\beta\longrightarrow\alpha aMa\beta \\
\qquad \alpha M\beta\longrightarrow\alpha bMb\beta \\
\qquad \alpha M\beta\longrightarrow\alpha aa\beta \\
\qquad \alpha M\beta\longrightarrow\alpha bb\beta
\end{cases}
$$

其中，α和β是 $(A\cup B)^*$上的任意的符号串。

序列M，aMa，$aaMaa$，$aabMbaa$是$aabMbaa$的一个推导，但不是扩展公理系统(3) 的一个证明，因为$aabMbaa$中还含有辅助字母表中的符号，它还不是定理。

序列M，aMa，$aaMaa$，$aabbaa$是扩展公理系统(3)的一个证明。

如果两个系统具有同样的证明的集合，则这两个系统等价。(3) 中的扩展公理系统与(2)中的公理系统是等价的。

挪威数学家图厄（Arel Thue）给扩展公理系统加上一定的限制，提出了半图厄系统，定义如下：

如果扩展公理系统（A，B，S，P）中的每个生成式格式都是双项的，并且具有形式

$$\alpha x\beta\rightarrow\alpha y\beta,$$

则这个扩展公理系统叫做半图厄系统。其中，x和y是 $(A\cup B)^*$上的符号串，α和β是变量，它们取 $(A\cup B)^*$上的符号串为其值。

在半图厄系统中，使用任何生成式所造成的变化，只限于用一个固定的符号串来替换另一个固定的符号串。显而易见，(2) 和(3) 两个系统都是半图厄系统，它们中的生成式全都是双项的。

由于半图厄系统中的一切生成式都是双项的，所以，我们可以把推导的定义限制得更窄一些。

给定一个半图厄系统（A，B，S，P），符号串y_1，y_2，…，y_m的线性有序序列叫做y_m的一个推导，当且仅当

1.y_1是一个公理；

而且，

2.除y_1之外，其余的每一个符号串都是从它直接前面的符号串使用P中的一个生成式导出的。

半图厄系统中，定理与证明的含义与扩展公理系统中的含义相同。

例如，有如下的半图厄系统（A，B，S，P），其中，

$A = \{C, D, E, F, G, H\}$

$B = \{a\}$

$S = \{HFGa\}$

P:

$FG \to DGaa$	①
$FD \to DF$	②
$HD \to HC$	③
$CD \to FC$	④
$CG \to FFGa$	⑤
$HF \to E$	⑥
$EF \to E$	⑦
$EG \to E$	⑧
$Ea \to a$	⑨

我们给出半图厄系统的符号串$aaaaa$的推导如下：

$HFGa$	公理
$HDGaaa$	①
$HCGaaa$	③
$HFFGaaaa$	⑤
$EFGaaaa$	⑥
$EGaaaa$	⑦
$Eaaaa$	⑧
$aaaaa$	⑨

乔姆斯基把文法定义为四元组$G=(V_N, V_T, S, P)$，与半图厄系统相比较，可以看出，文法中的V_N和V_T，分别相当于半图厄系统中的辅助字母表和基本字母表，文法中的重写规则P，相当于半图厄系统中的生成式，文法中的初始符号S，相当于半图厄系统中的公理，由文法推导出的终极符号串，相当于半图厄系统中的定理。在这个意义上，可以说，形式文法在实际上乃是一个公理系统。乔姆斯基的文法理论，不过是数学中的公理系统理论在

自然语言分析中的应用而已，语言就是由文法这一公理系统从初始符号S出发推出的无限定理的集合。文法的规则是有限的，终极符号和非终极符号的数目也是有限的，可是，由于语言符号具有递归性，文法这一公理系统就能够根据有限的符号，通过有限的重写规则，递归地推导出无限的语言来。采用这样的公理化方法，利用语言符号的递归性，乔姆斯基出色地实现了洪堡德提出的“有限手段的无限运用”这一原则。

由于语言的生成过程可通过公理系统这一形式化的手段得到严格的描述，所以，乔姆斯基的形式语言理论，在计算机程序语言的设计中，在自然语言信息处理的研究中（如机器翻译、人机对话），得到了广泛的应用，并且取得了令人满意的效果。

然而，我们知道，任何公理系统都是一个封闭的自足的系统，作为公理系统的文法，甚至象上下文无关文法这样比较适于描写自然语言的文法，它的描写界限，也仅只能局限于一个句子之内，只能说明一个句子本身的生成过程，它在公理系统内是自封闭的。由于公理系统的这种自封闭性质，使得被上下文无关文法所描写的一个句子，不可能与其它的句子发生联系。

例如，汉语中“张三借李四一本书”这个句子有歧义，要了解这个句子的实际含义，必须知道这个句子是从什么样的句子转换来的，由于来自不同的句子，它才会得到不同的语义解释；如果这个句子是从“张三借给李四一本书”转换来的，它可以得到一种语义解释，即：张三的书借给了李四；如果这个句子是从“李四借给张三一本书”转换来的，它又可以得到另一种不同的解释，即：李四的书借给了张三。这两种不同的含义，单从“张三借李四一本书”这个句子本身是解释不清楚的。因为作为公理系统的文法是自封闭的，它解决不了一个句子与其它句子的关系问题，这正是作为公理系统的文法的局限性，要解决这个问题，应该跳出公理系统的框架，到公理系统之外去研究句子与句子之间的关系。

乔姆斯基后来提出的转换语法，正是为了摆脱公理系统的束缚，从而在更为广泛的语言平面上，对自然语言进行描述和解释，所以，转换语法是对生成语法公理化方法的挑战。

第五章

语言符号的层次性与图论

第1节　语言符号的层次性

索绪尔认为语言符号具有线条性，它是只在时间上展开的，因而体现为一个长度，而这长度只能在一个向度上测定，它是一条直线。我们在“绪言”中已经指出，索绪尔的这个论断受到了语言学新的研究成果的严重挑战，英国语言学家弗斯的“跨音段论”就证明了，语言符号并不是线条性的东西，而是立体性的东西。

弗斯的“跨音段论”只限于音位学方面。其实，在语言的其它方面，语言结构也不仅仅只是线条性的，而是立体性的。所谓立体性，就是具有分层结构，也就是层次性。

语言符号的层次性，在句子结构方面表现得特别明显。

美国描写语言学派的语言学家们早就指出，英语的“The old men and women stayed at home”（年老的男人和女人留在家里）这句话是有歧义的。如果我们把这一句话说给一些人听，很可能有的听话人会认为它的意思是“年老的男人和所有的女人

(不论年龄大小)留在家里”，另一些听话人会认为它的意思是“所有年老的男人和所有年老的女人留在家里”，还有的听话人干脆不能作出决定，这与看图人看到一个中空的立方体图形时的情形很相似。看图的人可以差不多随心所欲地来看，先把立方体的某一角看作最靠近自己，然后又把另一个角看作最靠近自己。因此，一般人看到下面的A、B、C三个图形时，很容易看出B跟A和C都不同，但是，他会感到B有时更象A，有时又更象C。这样，B就是模棱两可的，或者说，B是有歧义的。如图5.1.1所示。

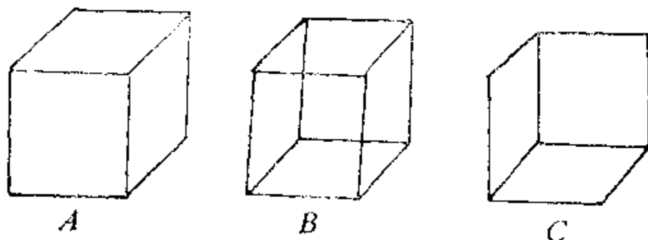


图5.1.1　B既象A又象C

事实上，“old men and women”这个名词短语根据意义的不同有两种不同的层次结构，如果注意到层次的不同，那么，这种意义上模棱两可的情况就可以得到解释。

一种层次结构是

old men and women
|______|　|__________|

这时，这个名词短语的意义是：“年老的男人和所有的女人”。

另一种层次结构是

old men and women
|__|　|______________|

这时，这个名词短语的意义是：“所有年老的男人和所有年老的女人”。

一般地说，如果要判断两个语言片段$A = a_1 a_2 \cdots a_n$和$B = b_1 b_2 \cdots b_m$是否具有同一性，至少应该满足3个条件：

1．A和B中对应的词形相同，词数相同。即有$a_1 = b_1$，$a_2 = b_2$，…，$a_n = b_m$，且$n = m$。

2．A和B中的词序相同。即：如果有$a_1 \prec a_2$，…，$a_{n-1} \prec a_n$，那么，则有$b_1 \prec b_2$，…，$b_{m-1} \prec b_m$。其中，"$\prec$"表示"前于关系"。

3．A和B中各个词之间的层次结构相同。

索绪尔主张语言符号的线条性，只看到了第1条和第2条，而没有看到第3条，这是他的局限性，今天，我们看到了第3条，发现了语言符号的层次性，应该说是一个很大的进步。

汉语有一个笑话也十分生动地说明了语言符号的层次性。客人希望留宿，先写下六个字："下雨天，留客天"。主人添上四个字，稍为改动一下层次，变成一条逐客令："下雨，天留客；天留，人不留"。客人把这一句的层次再加改动，又成了另外的意思："下雨天，留客天。留人不？留"。同样十个语素，同样的排列顺序，只是层次组合不同，意思就完全两样。

算命先生给人判断："父在母先亡"。可以作模棱两可的解释："父在，母先亡"，意思是母死父在；"父在母先，亡"，意思是父死母在。同样五个语素，同样的排列顺序，由于层次组合不同，意思就大相迳庭。这样，算命先生就成了永远灵验的铁嘴先生了。

上面的两个汉语例子多少有些人为性，似乎象是在做文字游戏。其实，这样由于层次组合不同而产生歧义的情况，在汉语里是普遍存在的。我国著名语言学家朱德熙先生早在1962年发表的《论句法结构》一文中①，就指出了因层次组合不同而产生的歧义，他举的例子是"咬死了猎人的狗"，可以有两种解释：一种是："咬死了猎人的狗"意思是猎人的狗被咬死了；一种是："咬死了猎人的狗"，意思是狗把猎人咬死了。在句法结构上，两种解释也各不相同：前者是述宾结构，后者是偏正结构。

①朱德熙，《论句法结构》，（《中国语文》），1962年，8—9月。

我们还可以在汉语的日常语言中找出许多这种因层次组合不同而造成歧义的例子。

例1．"发现了敌人的哨兵"，

一种层次结构是："发现了敌人的哨兵"，

这是一个述宾结构；

一种层次结构是："发现了敌人的哨兵"，

这是一个偏正结构。

例2．"热爱人民的总理"

一种层次结构是："热爱人民的总理"，这是一个述宾结构；

一种层次结构是："热爱人民的总理" 这是一个偏正结构。

例3．"哥哥和弟弟的朋友"

一种层次结构是："哥哥和弟弟的朋友"，这是一个联合结构；

一种层次结构是："哥哥和弟弟的朋友"，这是一个偏正结构。

例4．"雨来的小朋友铁头和小黑"

一种层次结构是："雨来的小朋友铁头和小黑" 这是一个同位结构。

一种层次结构是："雨来的小朋友铁头和小黑" 这是一个联合结构。

这些例子说明了，在日常语言的线性符号序列的内部，还隐藏着一个非线性的层次结构。我们举的例子比较简单，层次结构不十分复杂。而我们使用的句子一般都不会这样简单，有的句子的层次可以分为若干层，这时，就要用树形图（tree graph）才能把这种层次清楚地表示出来了。

例如，英语中They are flying planes 这个句子有两个不同的意思，这两个不同的意思是由于这个句子的线性序列的表层之下，隐藏着两个层次不同的树形图而造成的。

当其意思为“它们是正在飞的飞机”时，其树形图为5.1.2。

图5.1.2中，S表示句子，NP表示名词短语，VP表示动词短语，V表示变位动词，Ving表示词尾为-ing的动词，N表示名词。这时，flying是planes的定语，flying planes构成一个名词词组，are是系词。

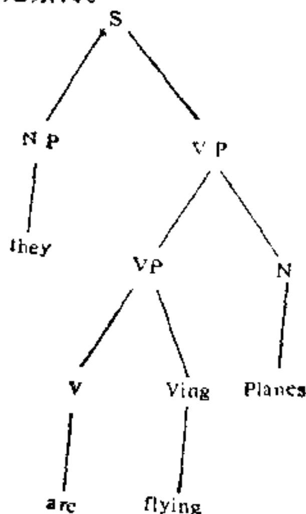


图5.1.2 树形图

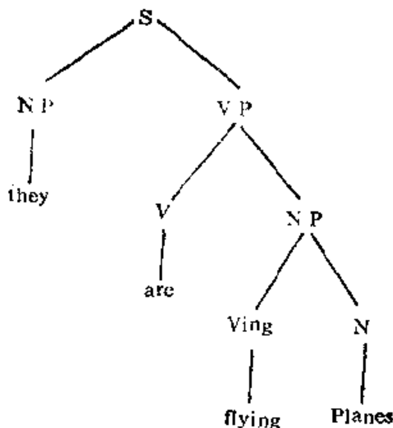


图5.1.3 树形图

当其意思是“他们正在驾驶飞机”时，其树形图为5.1.3。

其中，are和flying构成动词的现在进行时，planes作动词的直接宾语。

任何一个句子的线性序列的表层之下，都隐藏着一个层次分明的树形图。当一个句子的线性序列之下隐藏着两个或两个以上的树形图时，这个句子就会产生歧义，就会得到不同的解释。

我们前面所举的那些因层次不同而造成歧义的那些简单的例子，也是会有这样的树形图的。例如，英语“old men ald women”这个歧义的名词短语，所隐藏的两个不同的树形图如下：

当其意思是“年老的男人和所有的女人”时其树形图为5.1.4。

当其意思是“年老的男人和年老的女人”时，其树形图为

5.1.5。

在图5.1.4和5.1.5中，ADJ表示形容词，CONJ表示连接词，其它符号的含义如前所述。

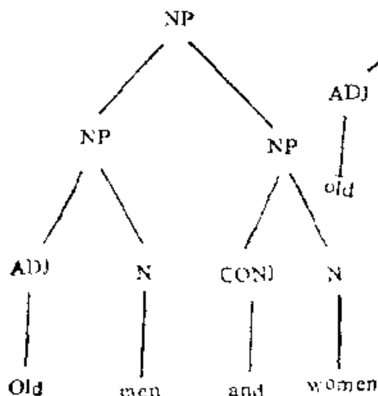


图5.1.4 树形图

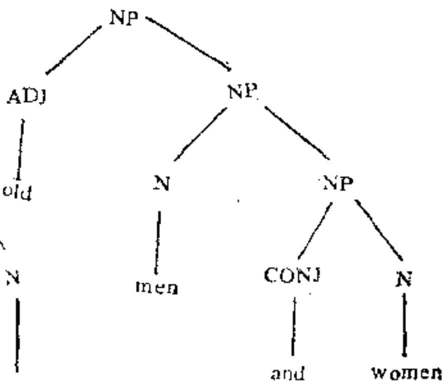


图5.1.5 树形图

由此可见，树形图是表示语言符号的层次性的一种很直观的形式，它可以把语言在句法结构上的层次差异揭示无余。

第2节 树 形 图

树形图是不包含回路的连通图。由于树形图可以直观地描述语言的层次结构，所以，语言研究便与数学中的图论发生了联系。

从直观上说来，树形图可以表示关于句子的句法结构的三个方面的信息：

1.句子中各成分的语法类型；

2.句子中各成分从左到右的线性顺序；

3.句子各成分的层次。

例如，在图5.2.1所示的树形图中，PRO表示代词，NUM表

示数量短语，CAR表示基数词，QTF表示量词，其它符号含义与前述相同。

从图5.2.1中可以看出，“我”是代词，“妹妹”是名词，“看见”是动词，“一”是基数词，“只”是量词，“猫”是名词，句中各成分的语法类型是很清楚的。从图中还可看出，在句子这个成分中，名词短语前于动词短语，在名词短语中，代词前于名词，…，句子中各成分从左到右的线性顺序也是清楚的。图中还可看出，标有S的最大成分由NP和VP两个成分组成，而NP又由PRO和N组成，VP由V和NP组成，后一个NP又由NUM和N组成，NUM由CAR和QTF组成，句子中各成分的层次组合关系也是很清楚的。

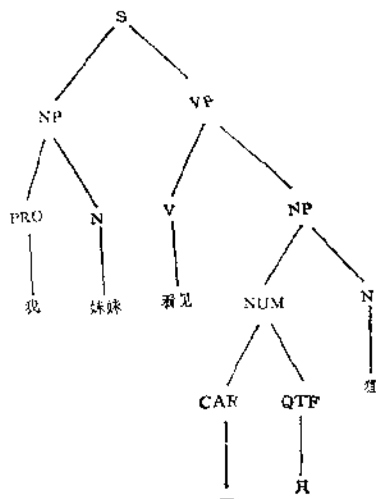


图5.2.1　树形图

树形图本身由结(node)和连接结的枝(branch)组成。每一个结有一个标记(label)，这个标记是从语法范畴(如S，NP，VP，N，V，…等)和符号串元素(如“我”、“妹妹”，…等)的有限集合中选出的。我们习惯上把树形图看成是在书页上竖直正立的，标有S的结在顶上，标有符号串元素的结在底处，在树的竖直方向上，枝总是从较高的结向较低的结延伸。

如果枝用箭头而不是用线段画出，那么，结与结之间的相对的竖直位置就成了树形图的无关特征，例如，图5.2.2中的四个图形表示的是同一个树形图。

这说明了，树形图是有向图，我们用树形图表示的信息特征是按习惯画出的。

树形图中各个结点之间，有两种关系值得注意：一种是支配

关系，一种是前于关系。

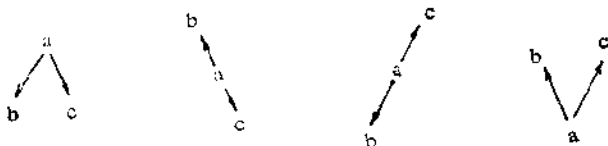


图5.2.2　枝为箭头的树

如果在树形图中，从结x到结y有一系列的枝把它们连接起来，而且从x到y的所有的枝有着同一方向，那么，我们就说结x支配结y，记为$D(x, y)$。例如，在图5.2.1的树形图中，标有VP的结支配标有CAR的结，因为连接结VP与结CAR的枝都一律从比较高的结VP，逐次通过结NP和结NUM，最后降到较低的结CAR。但是，标有VP的结不支配标有"我"的结，因为连接这两个结的枝首先要从结VP升到结S，再从结S通过结NP及结PRO降到结"我"。当x支配y时，y就叫做x的后裔(descendant)。

如果结x与结y是相异的，x支配y，而且，x与y之间没有另一个相异的结，那么，就说，x直接支配y。在图5.2.1中的树形图中，标有VP的结直接支配标有V的结，但不直接支配标有CAR的结。当结x直接支配结y时，结y就叫做结x的直接后裔或儿子。被同一个结直接支配的相异的结，叫做兄弟。图5.2.1中，标有VP的结有两个直接后裔，即标有V的结和右边标有NP的结，V和NP这两个结是兄弟。支配关系中不被任何其它结支配的结，叫做根。图5.2.1中，标有S的结就是根。被其它结支配而不支配任何其它结的结，叫做叶，图5.2.1中，标有终极标记的结"我"、"妹妹"、"看见"、…等等都是叶。按习惯，树形图是从上到下画出的，所以，根总是在顶部，叶总是在底部。

对于每一个合格的树形图，应满足单根条件：在每一个树形图中，恰好只有一个结是支配每一个结的，这个结就是根。

这个条件可写为：

$$(\exists x \in N)(\forall y \in N)\ (x, y) \in D$$

其中，N表示结的有限集合，$(x, y)\in D$表示x支配y。

树形图中的两个结，只有当它们之间没有支配关系的时候，才能在从左到右的方向上排序。这时，这两个结之间，就存在前于关系左边的结前于右边的结。在图5.2.1中，标有“我”的结前于标有VP的结以及所有被VP支配的结，因为结VP与结“我”之间不存在支配关系。但是，标有“我”的结不能前于支配它的结NP和结N，可见，支配关系同从左到右的前于关系是相互排斥的。

给定一个树形图，所有使得x前于y的有序对(x, y)，构成了这个树形图的一个前于关系，记为$P(x, y)$。

为了保证前于关系和支配关系没有共同的有序对，树形图应该满足如下的互斥条件：

在任何树形图中，对于任何的两个结x与y，x与y处于前于关系$P(x, y)$中，即或者$(x, y)\in P$，或者$(y, x)\in P$，当且仅当x与y不处于支配关系中，即$(x, y)\notin D$且$(y, x)\notin D$。

形式地说，互斥条件可写为：

$(\forall x, y\in N)\{[(x, y)\in P\vee(y, x)\in P\Longleftrightarrow[(x, y)\notin D\wedge(y, x)\notin D]\}$

对于任何的结x，$(x, x)\in D$，所以，$(x, x)\notin P$。

另外，在树形图中，还应该排除某些病态的结构：

1.树形图中不存在一个以上的枝进入的结，也就是说，每一个结最多只有一个直接支配它的结。

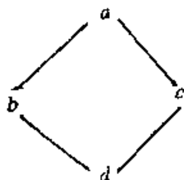


图5.2.3　病态结构之一

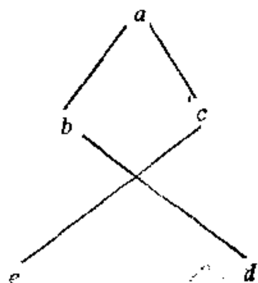


图5.2.4　病态结构之二

在图5.2.3.中，结d有两个直接支配它的结b和c，因此，这个图不是树形图。

2.树形图中不容许相互交叉的枝。

在图5.2.4中，连接b和d的枝与连接c和e的枝交叉了，因此，这个图也不是树形图。

为了排除这两种病态结构，人们提出了如下的非交条件：

在任何的树形图中，对于任何结x与y，如果x前于y，则由x支配的所有的结前于由y支配的所有的结。

形式地说，非交条件可写为：

$$(\forall w, x, y, z \in N)\{[(w, x) \in P \wedge (w, y) \in D \wedge (x, z) \in D] \Longrightarrow (y, z) \in P\}$$

在图5.2.3中，因为b前于c，b支配d，c也支配d，所以，d应该前于d，但这是不可能的，因而该图是一个病态结构。在图5.2.4中，因为b前于c，b支配d，c支配e，根据非交条件，d应该前于e，但实际上与此相反，所以，该图也是一个病态结构。

在图5.2.1中，每一个结有一个相应的标记，我们用标记函数L来表示结与标记之间的配对情况。这个标记函数的定义域是树形图中结点的集合，而其值域是有限的语法范畴和符号串元素的集合。

语法范畴S，NP，VP等，适于描写各种自然语言，它们对于各种自然语言是通用的。而符号串元素则因语言的不同而不同，数量又很多，因此，我们把标记函数L分为L_1和L_2两部分：L_1把叶的结点映入符号串元素F，L_2把非叶的结点映入语法范畴G，并且，G与F是不相交的。

根据上面所说明的各种关于树形图的性质，我们给出树形图的形式定义如下：

树形图可定义为一个五元组，用T表示：

$$T=(N, Q, D, P, L)$$

其中，N是一个有限集，即结的集合；

Q是一个有限集，即标记的集合；

D是在NXN上的支配关系；

P是在NXN上的前于关系，

L是N到Q的一个函数，即标记函数。

在这个五元组中，下列条件成立：

1.$(\exists x \in N)(\forall y \in N)(x, y) \in D$，即单根条件；

2.$(\forall x, y \in N)\{[(x, y) \in P \vee (y, x) \in P] \Longleftrightarrow [(x,y) \notin D \wedge (y, x) \notin D]\}$，即互斥条件。

3.$(\forall w, x, y, z \in N)\{[(w, x) \in P \wedge (w, y) \in D \wedge (x, z) \in D] \Longrightarrow (y, z) \in P\}$，即非交条件。

在这样的树形图中，对于N中的每一个结点，只有Q中的一个标记与之对应，所以，L是一个单值标记函数。

下面，我们来介绍三个对于刻画树形图的形式特性用得着的概念——属于、句友和统率。

1.属于：

给定一个树形图$T=(N, Q, D, P, L)$，结点x属于结点y，当且仅当

① $x \neq y$；

② $(y, x) \in D$；

③ $(y, S) \in L$；

④ $\neg(\exists w \in N)[(w, S) \in L \wedge w \neq y \wedge w \neq x \wedge (y, w) \in D \wedge (w, x) \in D]$

这个定义的②、③说明，x所属于的那个结y的标记为S，且支配x；④说明不允许任何的结S处于x与y之间并有支配关系；①排除了x结属于自身的情况。

结x属于y，记为$(x, y) \in B$。

这种情况，从图5.2.5的树形图中可以看出来。

结点“西山”属于带圈的结点S，因为这个带圈的S是支配“西山”且与“西山”离得最近的S。“西山”不属于根S，因为在根S与

“西山”之间，有带圈的S与它们处于支配关系之中。

“属于”这个概念可以把从属句中的各个成分与主句中的成分区别开来。

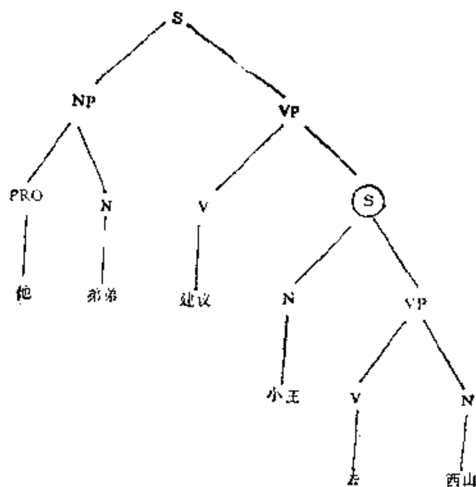


图5.2.5　带从句的树形图

2.句友：

给定一个树形图$T=(N, Q, D, P, L)$，结x与y是句友，当且仅当

$$(x, y)\notin D\wedge(y, x)\notin D\wedge\exists z\in N)[(x, z)\in B\wedge(y, z)\in B]$$

如果x与y是句友，则二者彼此互不支配，且都属于同一个结z。

在图5.2.5中，“小王”与“西山”是句友，因为这两个结彼此互不支配，而且都属于带圈的结S；“弟弟”和“西山”不是句友，因为它们不属于同一个结。

在简单句中，句友这一概念，可把句子内的成分与句子外的成分区别开来；在复合句中，句友这一概念，可把各分句中的成分区别开来。

3.统率：

给定一个树形图$T=(N, Q, D, P, L)$，结x统率结y，当且仅当：

$(x, y)\notin D\wedge(y, x)\notin D\wedge\exists z\in N[(x,z)\in B\wedge(z,y)\in D]$。

如果x统率y，则x与y彼此互不支配，且x属于一个支配y的结z。

在图5.2.5中，"弟弟"统率"西山"，因为二者互不支配，而"弟弟"属于根S，这个S支配"西山"；"西山"不统率"弟弟"，因为"西山"所属于的结是带圈的结S，这个结不支配"弟弟"。另外，我们还可以看出，在从句"小王去西山"中，"小王"统率"西山"，"西山"也统率"小王"。

统率这个概念，在主从复合句中，反映了主句中的成分对从句中各成分的统率作用，在简单句或同一分句内，它可以把句内成分和句外成分区别开来。

可见，"属于"、"句友"和"统率"这三个概念，对于句法结构的数学描述是大有好处的。

树形图与上下文无关文法有着密切的联系。我们可以用树形图来形象地表示上下文无关文法，它们之间的联系可通过下述方法来建立：

设$G=(V_N, V_T, S, P)$是上下文无关文法，其重写规则的形式是

$$A\longrightarrow\omega$$

其中，A是单个的非终极符号，ω是异于$\varnothing$的符号串，即有

$$|A|=1\leqslant|\omega|,$$

如果有某个树形图满足下列条件，它就是上下文无关文法G的推导树：

①每一个结有一个标记，这个标记是V中的符号；

②根的标记是S；

③如果结n至少有一个异于其本身的后裔，并有标记A，那么，A必定是V_N中的符号；

④如果结n_1，n_2，…n，n_k是结的直接后裔，从左向右排列，其标记分别为A_1，A_2，…A_k，那么。

$$A \longrightarrow A_1 A_2 \cdots A_k$$

必定是P中的重写规则。

例如，我们来考虑这样的文法。

$G=(V_N, V_T, S, P)$

$V_N=\{A, S\}$

$V_T=\{a, b\}$

$S=\{S\}$

P：$S \longrightarrow aAS$

$A \longrightarrow SbA$

$S \longrightarrow a$

$A \longrightarrow ba$

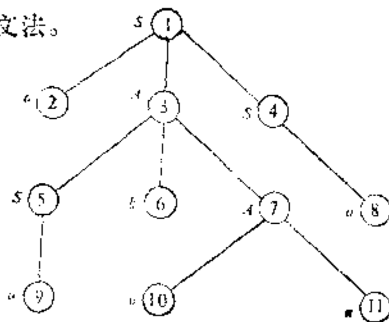


图5.2.6　上下文无关文法的推导树

这个文法的四个重写规则，左边都是单个的非终极符号S或者A，右边都是异于$\varnothing$的符号串，因而它是一个上下文无关文法。

现在，我们画出这个文法的推导树。为了便于说明，我们用圆圈表示结，并把结编上号码，把标记注在结的旁边，边的方向都假定是直接向下的。这个推导树见图5.2.6。

从这个树形图中可以看出：1，3，4，5，7等结都有直接后裔。结1是根，它的标记是S，其直接后裔从左算起为a，A和S，因而$S \longrightarrow aAS$是重写规则。结3的标记为A，其直接后裔的标记从左算起为S，b，A，因而$A \longrightarrow SbA$是重写规则。结4和结5的标记为S，它们每一个的直接后裔的标记为a，因而$S \longrightarrow a$是重写规则。结7的标记为A，其直接后裔从左算起为b和a，因而$A \longrightarrow ba$也是重写规则，由此可见，刚才画出的文法G的推导树满足推导树所要求的各个条件。

如果从左到右读推导树中各个叶的标记，就可以得到一个终极符号串，这个终极符号串叫做推导树的结果。可以证明，如果a是上下文无关文法$G=(V_N, V_T, S, P)$的结果，则有

$$S \overset{*}{\underset{G}{\Longrightarrow}} \alpha$$

例如，在上述推导树中，各个叶从左到右的编号为2，9,6，10，11和8，它们的标记分别是a，a，b，b，a和a,则推导树的结果$\alpha = aabbaa$，因此，

$$S \overset{*}{\underset{G}{\Longrightarrow}} aabbaa$$

在上面的推导树中，为了说明方便，我们给每个结编了号。在一般情况下，我们并没有必要给结编号，而在结上直接写上其标记。例如，上面的推导树的习惯画法如图5.2.7所示。

其推导过程为

$S \Longrightarrow a\mathrm{AS} \Longrightarrow a\mathrm{S}b\mathrm{AS} \Longrightarrow aab\mathrm{AS} \Longrightarrow aabba\mathrm{S} \Longrightarrow aabbaa$。

乔姆斯基证明了，任何的上下文无关语言，均可由重写规则为

$$A \longrightarrow BC \quad 或 \quad A \rightarrow a$$

的上下文无关文法生成，其中，A，B，$C \in V_N$，$a \in V_T$。这种规则叫做乔姆斯基范式。

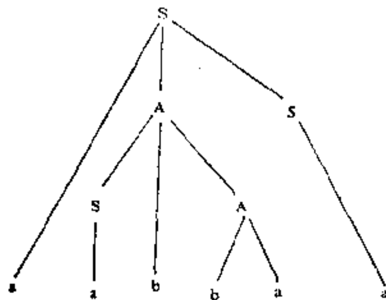


图5.2.7　推导树的习惯画法

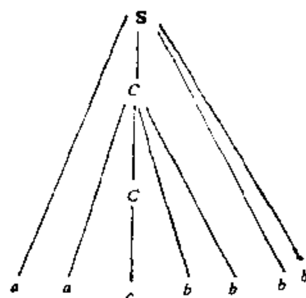


图5.2.8　生成符号串$aacbbbb$的推导树

利用乔姆斯基范式，可以把任何的上下文无关文法的推导树简化为二元形式。

例如，上下文无关语言$\{a^n cb^{2n}\}$的文法的重写规则为

$$S \longrightarrow aCbb$$

$$C \to aCbb$$

$$C \to c$$

如果要生成符号串$aacbbbb$，其推导树如图2.5.8。

现在我们把这个文法的三个重写规则改写为乔姆斯基范式。

在这三个规则中，$C\to c$是符合乔姆斯基范式要求的，不必再变换。我们先把$S\to aCbb$及$C\to aCbb$ 的右边换为非终极符号，用$S\to ACBB$及$A\to a$，$B\to b$来替换$S\to aCbb$，用$C\to ACBB$及$A\to a$，$B\to b$来替换$C\to aCbb$。然后，再把$S\to ACBB$，$C\to ACBB$ 的右边换成二元形式，用$S\to DE$，$D\to AC$及$E\to BB$来替换$S\to ACBB$，用$C\to DE$，$D\to AC$及$E\to BB$来替换 $C\to ACBB$。这样，便得到了符合乔姆斯基范式要求的文法的重写规则：

$$S\to DE$$

$$D\to AC$$

$$E\to BB$$

$$C\to DE$$

$$A\to a$$

$$B\to b$$

$$C\to c$$

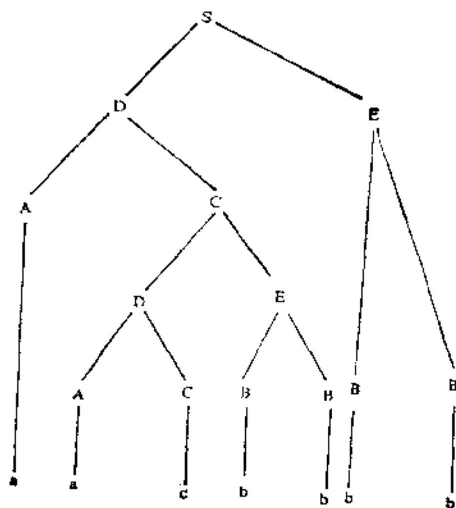


图5.2.9

用乔姆斯基范式，可将符号串$aacbbbb$ 的推导树简化为二元形式，这种二元形式的推导树又叫二叉树。如图5.2.9所示。

乔姆斯基范式的重写规则及推导树都具有二元形式，这就为

自然语言的形式描写提供了数学模型。

自然语言中的句法结构一般都是二分的，因而一般都具有二元形式。汉语中由实词和实词性词组组合成的句法结构绝大部分都是二元形式的结构。例如：

1.主谓结构：小王工作

2.偏正结构：优秀学生

3.述宾结构：克服困难

4.述补结构：洗干净

5.联合结构：水果蔬菜

6.复谓结构：去看电影

事实上，语言学中正是采用二分法来分析句子的。二分法就是所谓的层次分析法，这种分析法认为，一个复杂的语言形式，不能一下子就把它分析为若干个词，而要按下面的步骤逐层地进行分析：

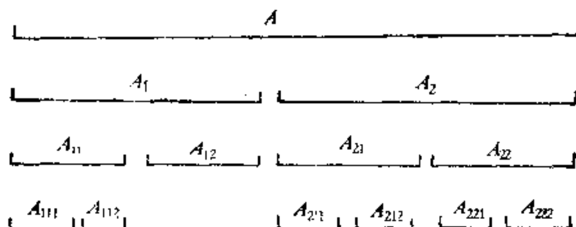


图5.2.10 层次分析法示意图

从图5.2.11中可以看出，我们不是把A一下子就分成A_{111}，A_{112}，A_{211}，A_{212}，A_{221}，A_{222}这七个部分的，而是先把A分成A_1和A_2两部分，然后再把A_1分成A_{11}和A_{12}两部分，把A_2分成A_{21}和A_{22}两部分，又把A_{11}分成A_{111}和A_{112}两部分，把A_{21}分成A_{211}和A_{212}两个部分，…等等，这样分析下去，一直分析到单词为止。人们通常把A_1和A_2叫做A的直接成分，把A_{11}和A_{12}叫做A_1的直

接成分，把A_{111}和A_{112}叫做A_{11}的直接成分，……，等等。这种顺次找出语言格式的直接成分的方法，叫做直接成分分析法或层次分析法，因此，在语言学界，又有人把乔姆斯基的上下文无关文法叫做直接成分语法或短语结构语法。短语结构语法成为一个广泛使用的语言学术语，在机器翻译和计算语言学中得到了进一步的检验和研究。

由于乔姆斯基范式反映了自然语言的二分特性，因而通过乔姆斯基范式这一重要工具，短语结构语法成为了自然语言研究中的一种最基本的语法。

事实上，不少语言学家在他们描写自然语言的研究中，早已认识到了自然语言的这种二分特性。

我国语言学家在《马氏文通》中提出了"两端两语说"，他指出："盖意非两端不明，而句非两语不成"。美国语言学家奈达(E. A. Nida)在《形态学》中指出："根据经验，我们发现语言结构倾向于二分"①。美国语言学家福里斯(C.C.Fries)在《英语结构》一书中，更是明确地提出了二分的观点，他指出："在英语里，一个结构层次通常只有两个成分。当然，每一个成分都可以由好几个单位组成，不过在同一层次上，结构的直接成分通常只有两个"②。由此可见，图论给语言学家提供的二叉树，确实是语言研究的一个有力手段。

---

① E. A. Nida, *Morphology*, *University of Michigan*, 1949, 第91—93页。

② C. C. Fries,《英语结构》(中译本)，商务印书馆，第264页。

第六章

语言符号的非单元性与复杂特征的运算

第1节　语言符号的非单元性

在第五章中我们说过，短语结构语法在图论中的描述形式是单标记的二叉树形图，这种树形图反映了自然语言的二分特性。但在具体的自然语言研究中，特别是在机器翻译等自然语言的计算机处理研究中，我们发现这种树形图有两个严重的缺陷：

第一，虽然自然语言的结构倾向于二分，但二分法并不是到处行得通的，特别是在汉语中，许多语法形式看来宜于采用多分法。例如

1.双宾语结构：

给弟弟一本书

2.兼语式结构：

请他做报告

3.多于两项的联合结构：

小张、小李和小王

4."状语+述语+宾语"结构：

努力学习数学

5.框形方位结构：

在工作中

在这些情况下，采用多分法的好处是：

①可以更加合理地解释语言现象。例如，"请他做报告"中，"他"作"请"的宾语，又作"做报告"的主语，一身而二任，如果采用二分法，在树形图上就会发生交叉现象，违反了树形图的"非交条件"，破坏了树形图的结构，采用多分法分析为"请|他|做报告"三部分，用一个三叉树形图来表示其结构，便不会发生交叉现象。"给弟弟一本书"中，"给"有两个宾语，采用多分法，用一个三叉树形图来表示其结构，其层次更为清楚。"努力学习数学"中，状语"努力"究竟是修饰述语"学习"，还是修饰述语+宾语"学习数学"，从语感上很难判别，一次就把它分为三部分，避免了解释上的困难。

②可以在自然语言自动处理中减少编制程序的工作量：一些长句子，如果采用二分法，层次会多到十层八层，计算机在处理这样的多层次的树形图时，需逐层进行，运算量很大，而采用多分法，大大减少了层次，提高了自然语言计算机处理的工作效率。

③可以抓住句子的主干，把句子的格局清楚地显示出来，便于检查和研究。

因此，我们采用多叉树形图来代替二叉树形图，而且把多叉树形图看成是一种普遍的树形图格式，把二叉树形图看成是多叉树形图的一种特殊情况。所谓"多叉"，可以是"三叉"、"四叉"，也可以是"二叉"、"一叉"，它是一种更为一般的形式，而"二叉"只不过是当"多叉"的"多"是"二"时的一种特殊情况罢了。

第二，单标记树形图的标记太简单，不宜于区分自然语言中

的歧义结构。

在短语结构语法的推导树$T=(N, Q, D, P, L)$中，L是从N到Q的标记函数。这种标记函数是单值标记函数，也就是说，对应于结点集合N中某一元素x，有标记集合Q中的一个元素y与之对应，这样的单值标记函数，可记为

$$L(x)=y$$

在短语结构语法中，标记一般是用词类或词组类型等非终极符号以及符号串元素等终极符号来描述的，在非叶结点上的标记是非终极符号，在叶结点上标记是终极符号。

这种单值标记函数表示的语言特征是十分有限的，因而会产生大量的歧义结构，形成大量不合语法的句子，这是单值标记函数的最大缺点，也是以单值标记函数为特征的短语结构语法的最大缺点。

自然语言的句子不能只用词类或词组类型等特征来描述，特别是在汉语中，句子各个成分的词组类型、句法功能、语义关系、逻辑关系之间，存在着极为错综复杂的关系，如果只使用词类或词组类型等简单特征，就难以区别各种歧义现象，这样，在汉语自动处理中，就达不到自动句法语义分析的目的。具体地说：

①汉语句子中的词组类型（或词类）与句法功能之间不存在简单的一一对应关系。

用短语结构语法的单值标记函数来分析英语句子时，对于树形图中的每一个结点，只给关于词组类型或词类的特征，如S，NP，VP，Ad_j，N，V等，一般不会碰到很大的困难。因为在英语中，一旦把S分解为NP和VP，那么，NP一般是主语，VP一般是谓语，形成一个主谓结构；一旦把VP分解为V和NP，那么，V一般是述语，NP一般是宾语，形成一个述宾结构；句子组成成分的词组类型和句法功能之间存在着比较简单的一一对应关系。当句子各个成分的句法功能关系确定之后，也就不难进一步确定这些成分之间的语义关系和逻辑关系，从而实现句子的句法分析和

语义分析。

但是，在汉语中，仅仅使用词组类型（或词类）这样的特征是远远不够的，因为汉语句子中的词组类型（或词类）与句法功能之间不存在简单的一一对应关系。一个NP加上一个VP,可以构成主谓结构（如"小王/咳嗽"),但也可以构成偏正结构，如"程序/设计"，"程序"是NP，不作主语而作定语,"设计"是VP，不作谓语而作被修饰的中心语。类似的例子还有"语言/学习"、"政治/工作"、"物理/考试"等，词组类型都是NP + VP,可是，不形成主谓结构，而形成偏正结构，在这种情况下，如果只用词组类型这样的简单特征NP + VP 就不能区别这种结构在句法功能上的歧义,而必须既使用词组类型特征,又使用句法功能特征,这样,我们在树形图的结点上，就不能采用单标记,而必须采用多标记了。

采用多标记，对于形成主谓结构的NP + VP，可描述为

$$\left[\begin{bmatrix} K = NP \\ CAT = N \\ FS = SUBJ \end{bmatrix} + \begin{bmatrix} K = VP \\ CAT = V \\ FS = PRED \end{bmatrix}\right]$$

式中，K表示词组类型特征，NP和VP都是K这个特征的值，它们形成一类标记；CAT表示词类特征，N和V都是CAT这个特征的值，它们又形成另一类标记；FS表示句法功能特征，SUBJ和PRED是FS这个特征的值，SUBJ表示主语，PRED表示谓语，它又形成一类新标记。这样，这一类结构的每一个结点上，就不再只有一个单标记，而是具有三个标记，形成多标记的结构。

对于形成偏正结构的NP + VP，可描述为

$$\left[\begin{bmatrix} K = NP \\ CAT = N \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = VP \\ CAT = V \\ SF = HEAC \end{bmatrix}\right]$$

式中，MODF表示定语，HEAD表示中心语，它们是SF这个特征

的值。

对于这两种词组类型相同而句法功能不同的结构，如果只用单标记的简单特征NP＋VP来描述，显然就不能反映它们在句法功能方面的差异，必须同时用词组类型特征和句法功能特征结合而成的多标记，才能准确地描述它们。

汉语中一个VP加上一个NP，可以形成述宾结构(如“学习/英语”)，但也可以形成偏正结构，如“出租/汽车”中，“出租”是VP，不作述语而作定语，“汽车”是NP，不作“出租”的宾语而作被“出租”修饰的中心语。类似的例子很多,如“研究/方法”、“学习/制度”、“开放/政策”等，词组类型都是VP＋NP，可是，不形成述宾结构，而形成偏正结构。在这种情况下，如果采用单标记的简单特征VP＋NP来描述，就会产生句法功能歧义,而必须采用多标记的方法来描述，既使用词组类型特征，又使用句法功能特征，才能把这种歧义区别开来。

对于形成述宾结构的VP＋NP，可描述为

$$\left[\begin{bmatrix} \text{K}=\text{VP} \\ \text{CAT}=\text{V} \\ \text{FS}=\text{PRED} \end{bmatrix} + \begin{bmatrix} \text{K}=\text{NP} \\ \text{CAT}=\text{N} \\ \text{FS}=\text{OBJE} \end{bmatrix}\right]$$

式中，PRED表示述语，OBJE表示宾语，它们都是句法功能特征SF的值。

对于形成偏正结构的VP＋NP，描述为

$$\left[\begin{bmatrix} \text{K}=\text{VP} \\ \text{CAT}=\text{V} \\ \text{FS}=\text{MODF} \end{bmatrix} + \begin{bmatrix} \text{K}=\text{NP} \\ \text{CAT}=\text{N} \\ \text{FS}=\text{HEAD} \end{bmatrix}\right]$$

式中，MODF表示定语，HEAD表示中心语，它们是句法功能特征FS的值。

对于这两种词组类型相同而句法功能不同的结构，如果只用

单标记的简单特征VP+NP来描述，显然也是不充分的，必须采用多标记的方法来描述。

②汉语句子中词组类型(或词类)和句法功能都相同的成分，它们与句中其它成分的语义关系还可能不同，句法功能和语义关系之间也不是简单地一一对应的。

同样是由NP和VP组成的主谓结构，其中作主语的NP的语义可以是施事者（如“小王/工作”中的“小王”），也可以是受事者（如“火车票/买了”中的“火车票”），还可以是工具（如“左手/拿纸，右手/拿笔”中的“左手”和“右手”）。因此，在汉语句子的自动处理中，仅仅知道了句子的组成成分的词组类型特征和句法功能还不够，为了区分歧义，还要再加上语义关系特征来描述，这样，就更需要采用多标记的方法了。

对于NP的语义关系为施事者、句法功能为主语的NP+VP，可描述为

$$\left[\begin{bmatrix} K=NP \\ CAT=N \\ FS=SUBJ \\ SM=AGENT \end{bmatrix} + \begin{bmatrix} K=VP \\ CAT=V \\ FS=PRED \end{bmatrix}\right]$$

其中，SM表示语义关系特征，AGENT表示施事者，它是语义关系特征SM的值。

对于NP的语义关系为受事者，句法功能为主语的NP+VP，可描述为

$$\left[\begin{bmatrix} K=NP \\ CAT=N \\ SF=SUBJ \\ SM=PATIENT \end{bmatrix} + \begin{bmatrix} K=VP \\ CAT=V \\ SF=PRED \end{bmatrix}\right]$$

其中，PATIENT表示受事者，它是语义关系特征SM的值。

对于NP的语义关系为工具、句法功能为主语的NP+VP，可描述为

$$\begin{bmatrix}\begin{bmatrix}K=NP\\CAT=N\\SF=SUBJ\\SM=INST\end{bmatrix} & + & \begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix}\end{bmatrix}$$

其中，INST表示工具，它也是语义关系特征SM的值。

同样是由VP和NP组成的述宾结构，其中，作宾语的NP的语义关系更是复杂多样。在英语中，作宾语的NP一般表示述语VP的受事者，但在汉语中，作宾语的NP在语义关系上可以是述语VP的受事者、范围、目的、结果、工具等等。

例如，动词“考”后面加上不同的NP，作宾语，这些宾语NP与述语“考”的语义关系极为复杂。在“考/学生”中，宾语“学生”是“考”的受事者；在“考/数学”中，宾语“数学”是“考”的范围；在“考/北大”中，宾语“北大”是“考”的目的；在“考/研究生”中，宾语“研究生”是“考”的结果(“考/研究生”在语义上是有歧义的，在一定的环境下，“研究生”可以是“考”的受事者，是被考的人)；在“考/一百分”中，宾语“一百分”也是“考”的结果。因此，在中文句子的分析中，仅仅有了词组类型特征和句法功能特征还是不够的，还必须再加上语义关系特征。

对于NP的语义关系为受事者，句法功能为宾语的VP+NP，可描述为

$$\begin{bmatrix}\begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix} & + & \begin{bmatrix}K=NP\\CAT=N\\SF=OBJE\\SM=PATIENT\end{bmatrix}\end{bmatrix}$$

其中，PATIENT表示受事者，它是语义关系特征SM的值。

对于NP的语义关系为范围、句法功能为宾语的VP＋NP，可描述为

$$\left[\begin{bmatrix}\text{K}=\text{VP}\\ \text{CAT}=\text{V}\\ \text{SF}=\text{PRED}\end{bmatrix}+\begin{bmatrix}\text{K}=\text{NP}\\ \text{CAT}=\text{N}\\ \text{SF}=\text{OBJE}\\ \text{SM}=\text{SCALE}\end{bmatrix}\right]$$

其中，SCALE表示范围，它是语义关系特征SM的值。

对于NP的语义关系为目的、句法功能为宾语的VP＋NP，可描述为

$$\left[\begin{bmatrix}\text{K}=\text{VP}\\ \text{CAT}=\text{V}\\ \text{SF}=\text{PRED}\end{bmatrix}+\begin{bmatrix}\text{K}=\text{NP}\\ \text{CAT}=\text{N}\\ \text{SF}=\text{OBJE}\\ \text{SM}=\text{GOAL}\end{bmatrix}\right]$$

其中，GOAL表示目的，它是语义关系特征SM的值。

对于NP的语义关系为结果、句法功能为宾语的VP＋NP，可描述为

$$\left[\begin{bmatrix}\text{K}=\text{VP}\\ \text{CAT}=\text{V}\\ \text{SF}=\text{PRED}\end{bmatrix}+\begin{bmatrix}\text{K}=\text{NP}\\ \text{CAT}=\text{N}\\ \text{SF}=\text{OBJE}\\ \text{SM}=\text{RESULT}\end{bmatrix}\right]$$

其中，RESULT表示结果，它是语义关系特征SM的值。

③汉语中单词所固有的语法特征和语义特征，对于判别词组结构的性质，往往有很大的参考价值，除了词组类型这样单标记的简单特征之外，再加上单词固有的语法特征和语义特征，采用多标记的方法来描述，就可以判断词组结构的性质。

在VP＋NP这样的词组类型结构中，如果VP的语法特征是不

及物动词，那么，VP的句法功能必为定语，NP的句法功能必为中心语。例如，"示踪程序"中，"示踪"为VP，是一个不及物动词，"程序"为NP，因为不及物动词不能带宾语，因此，"程序"不能为"示踪"的宾语，这时"示踪"是定语，"程序"是中心语。这种情况，可以表示为

$$\begin{bmatrix} K = VP \\ CAT = V \\ TRANS = IV \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} \begin{bmatrix} K = VP \\ CAT = V \\ TRANS = IV \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SF = HEAD \end{bmatrix} \end{bmatrix}$$

式中，TRANS表示动词的及物性，IV表示该动词的及物性为不及物，它是特征TRANS的一个值。

这个式子说明，在VP+NP中，当VP的及物性的值为不及物时，VP的句法功能为定语，NP的句法功能为中心语。

由此可以看出单词本身固有的语法特征对判断词组的句法功能的作用。

此外，单词本身固有的语义特征，对于判断词组的句法功能也有很大的作用。

在词组类型结构VP+NP中，当VP为及物动词，即它的及物性为及物时，词组的句法功能特征，就可以根据NP的语法特征来判别。一般地说，当VP为及物动词，NP为抽象名词，即NP的固有语义特征为"抽象物"时，或者当NP为类名词，即NP的固有语义特征为"类别名称"时，VP的句法功能为定语，NP的句法功能为中心语。例如，"训练/目的"这个词组中，"训练"为及物动词，"目的"为抽象名词，即"目的"的固有语义为"抽象物"，

因此，可判断“训练”的句法功能为定语，“目的”的句法功能为中心语。类似的例子还有：“生产/宗旨、培养/目标、发展/方向、管理/体制、进攻/计划”等。又如，“管理/人员”这个词组中，“管理”为及物动词，“人员”为类名词，即“人员”的固有语义为“类别名称”，因此，可判断“管理”为修饰语，“人员”为中心语。类似的例子还有：“采购/人员，进修/教师、领导/干部、评论/专家、革新/能手、主治/医生”等。

前一种情况可以表示为

$$\begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SEM = ABS \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \end{bmatrix} \\ SF = MODF \end{bmatrix} + \begin{bmatrix} \begin{bmatrix} K = NP \\ CAT = N \\ SEM = ABS \end{bmatrix} \\ SF = HEAD \end{bmatrix}$$

后一种情况可表示为

$$\begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SEM = SORT \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \end{bmatrix} \\ SF = MODF \end{bmatrix} + \begin{bmatrix} \begin{bmatrix} K = NP \\ CAT = N \\ SEM = SORT \end{bmatrix} \\ SF = HEAD \end{bmatrix}$$

式中，TV表示“及物”，它是特征TRANS的一个值，ABS表示“抽象物”，它是特征SEM的一个值，SORT表示“类别名称”，它是特

征SEM的另一个值。这里，特征SEM与前述的特征SM不一样，SEM是单词本身固有的语义特征，它不反映单词与单词之间的语义关系，SM是单词的语义关系特征，它反映的正是单词与单词之间的语义关系。

由此可见，在汉语句子的描述中，仅仅采用词类或词组类型这样的单值标记是远远不够的，必须再加上句法功能特征和语义关系特征，甚至还要加上单词本身固有的语法和语义特征，才有可能比较全面地表达句子中包含的语言信息，从而也才有可能成功地进行汉语句子的自动分析，建立与汉语有关的机器翻译系统或人机对话系统，所以，我们在汉语句子的自动分析中，必须对乔姆斯基短语结构语法进行修正，采用多标记的方法，把单标记的树形图变为多标记的树形图。

以上我们只是对这个问题作了初步的论述，而实际的语言现象往往比我们想象的还要复杂得多。汉语中施事者和受事者有时很难分辨，常常需要语境方面的背景知识才能判别。例如，在"小王/理发"这个NP+VP中，如果"小王"是理发师，那么，"小王"一般应该是施事者，他给别人理发；如果"小王"不是理发师，而是被理发的人，那么，"小王"就是受事者。"小王"究竟是施事者还是受事者，是由"小王"的身份这种背景知识来判别，单凭语言本身是难以分辨的。这时，我们在树形图结点上加的标记，就势必要扩大到语境特征的范围了。这类例子并不少见。在"小王/修车"、"小王/拔牙"、"小王/看病"等NP+VP中，"小王"究竟是施事者还是受事者，都要通过语境特征的分析，才能作出正确的判别。在这些情况下，就更需要采用多标记的方法了。

基于上述原因，我国学者于1981年对乔姆斯基的短语结构语法进行了重要的改进，提出了多值标记函数的概念，并用多值标记函数来代替短语结构语法的单值标记函数。

多值标记函数L可表示如下：

$$L(x)=\begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{Bmatrix}$$

采用这样的多值标记函数，树形图中的一个结点x,不再仅仅对应于一个标记，而是对应于若干个标记$\{y_1, y_2, \cdots\cdots, y_n\}$。在同一个结点上采用多个标记，大大地提高了树形图的标记功能，使得树形图的各个结点上，都能记录尽可能多的语法语义信息，除了记录短语结构语法所常采用的词类或词组类型信息之外，还记录单词本身固有的语法语义信息、单词之间或词组之间的句法功能信息、单词之间或词组之间的语义关系信息、单词之间或词组之间的逻辑关系信息。实践证明，这样的多值标记函数有效地克服了短语结构语法的缺陷，在自动句法语义分析方面，特别是在辨别歧义结构方面，是非常有效的。

我国学者在多值标记函数的基础上，进一步采用多叉树形图来代替二叉树形图，提出了"汉语句子的多叉多标记树形图分析法"，这种分析法又叫做中文信息处理的"多叉多标记树模型(Multiple—branched and Multiple—labeled Tree Model，简称MMT模型)①。

根据MMT模型，我国学者于1981年进行了汉—法/英/日/俄/德多语言机器翻译试验，建立了FAJRA系统，接着，于1985年又利用IBM—4341计算机，在VM/CMS操作系统下，进行了德—汉机器翻译试验和法—汉机器翻译试验，建立了GCAT德—汉机器翻译系统和FCAT法—汉机器翻译系统。实验结果令人满意。实践证明，建立在多值标记函数基础上的MMT模型是描述汉语句子结构的一个较好的模型。

---

①冯志伟，《汉语句子的多叉多标记树形图分析法》，(《人工智能学报》)，1983年，第2期。

我们之所以提出多值标记函数，是基于我们对语言符号的非单元性的认识。

单值标记函数只允许一个结点一个标记，结点被看成是一个不可分的单元，就象古代原子论中的原子一样。当我们采用这种单值标记函数来对自然语言作自动分析而感到左支右绌、进退维谷的时候，我们曾经想过，象这样单元性的结点，是不是也是可分的呢？它是不是也是有结构的呢？在现代物理学中关于“原子无限可分”的理论的启发下，我们把树形图中的一个个结点想象成有结构的、由多种特征组合而成的非单元体。这种关于语言符号非单元性的新认识，有如一盏明灯，引导我们摆脱了进退维谷的困境，走进了一马平川的坦途。当我们放弃单值标记函数而决定采用多值标记函数的时候，横亘在汉语句子自动分析中的重重困难都涣然冰释了。

事实证明了，语言符号并不是一个无结构的单元性符号，而是一个有结构的、由多种特征组合而成的非单元符号。我们上文中采用多标记分析过的那些语言现象，足以说明语言符号的这种非单元性。

语言符号的这种非单元性不仅存在于句子结构中，也存在于语音中。早在1938年，美国语言学家雅可布逊(R. Jakobson) 在比利时的根特城举行的第三届国际语音学会议上，就提出了能否以对分法为基础来分解元音、辅音等音位的问题。1951年，他在与范特 (M. Fant)、哈勒 (M. Halle)等语音学家合写的论文《语音分析初探》中，提出了对分法理论以及区别特征学说。他们认为，一切语音的音（无论元音或辅音）都是可分的，可以根据它们的生理的或声学的特性，用对分法分成一对对的“最小对立体”(minimun pairs)。例如，元音的舌位有“高—低”的对立，辅音的发音方法有“清—浊”的对立。他们把这些最小对立体归纳为十二对区别特征 (distinctive features)，并且指出，世界上各种语言都可以用这十二对区别特征加以描述。这样，过去一直认为

不可分的单元性的元音、辅音就变成由若干区别特征组合而成的、非单元性的结构体了。

这十二对区别特征是：

①元音性—非元音性：如a—p。

②辅音性—非辅音性：如p—a。

③鼻音性—口音性：如m—p，n—t。

④聚集性—分散性：如e—i。发e时，频谱中心能量集中；发i时，频谱中心能量分散。

⑤突发性—延续性：如p—f，b—v。

⑥粗糙性—圆润性：如s—θ。发s时，发音狭缝边缘粗糙；发θ时，发音狭缝边缘光滑。

⑦急停性—非急停性：如p'—p。发p'时，气流突然减弱；发p时，气流逐渐减弱。

⑧浊音性—清音性：如v—f，b—p。

⑨紧张性—松弛性：如k—g。发k时，语音有一定的稳定阶段，发音器官肌肉比较紧张；发g时，语音的稳定阶段较短，发音器官肌肉松弛。

⑩钝音性—锐音性：如m—n。发m时，频谱的重心在低频区；发n时，频谱的重心在高频区。

⑪降音性—平音性：如u—i。发u时，频谱中的高频成分比发i时降低或减弱。

⑫升音性—平音性：如d_j—d。发d_j时，频谱中的高频成分比发d时升高或加强，而且，发d_j时，舌部上抵硬颚，产生颚化作用。

这种区别特征理论已成为现代语音学进行音位分析的基础。任何一个音位，都可以用区别特征的集合来加以描述。如某一音位具有二项对立中的前项特征，记以“+”号，具有二项对立中的后项特征，记以“-”号。这样，便可作为一个矩阵表，作为对每一音位的区别特征集合的描述。

音位对分法理论已在语音自动识别和合成的研究中得到应用，证明是行之有效的。

雅可布逊曾提到，他之所以提出音位对分法理论，是受到了现代物理学的影响所致。他写道："语音学分析及其得出的、不能再行分解的音位特征的概念，同现代物理学的研究成果有惊人的相同之处，物理学也正表明，物质具有粒子状结构，因为它是由基本粒子构成的。"①

物理学中关于物质具有粒子状结构的观点，音位学中关于音位由十二对基本的区别特征组合而成的观点，句子自动分析中关于语言符号由多个特征组合而成的观点，它们之间是何等的相似！客观世界中存在着的这种相似现象，说明了这些现象之间是有内在联系的，认识事物之间的这种相似性，可以增进我们进行科学研究的才干，提高研究工作的自觉性和目的性。英国物理学家法拉第（M. Faraday）受到他的老师戴维（H. Davy）把化学能转化为电能，又把电能转化为化学能的可逆过程的启发，立志要把已发现的由电生磁现象逆转为由磁生电，经过九年的努力，终于实现了由磁生电的实验，建立了电磁感应学说的完整理论。正是这种对于事物之间存在相似性的信念，使我们提出了反映语言符号非单元性的"多标记"的概念，并进而建立了中文信息处理的MMT模型。

在汉语句子的自动分析中，我们是用非单元性的特征的组合来代替单元性的单值标记的。这种非单元性的特征组合，可通过"特征/值"系统来描述。这种"特征/值"系统，也就是汉语的多标记系统，它具体地说明了多值标记函数中的 $\{y_1, y_2, \cdots\cdots, y_n\}$ 究竟应取些什么样的标记。

我们在描述上面的汉语句子时，是采用若干个特征和它们的值来进行描述的。汉语的多标记系统包含若干个特征，而每一个

① R. Jakobson, *On the identification of phonemc entities*, TCLP, Vol. V, 1949, 第213页。

特征又包含若干个值，这种由特征和它们的值构成的描述系统，叫做“特征/值”系统。每种语言都有自己的“特征/值”系统。语言不同，它们的“特征/值”系统也不尽相同。

汉语的“特征/值”系统如下：

1.词类特征和它的值：

词类是描述汉语句子的一个重要特征，在短语结构语法中，词类是常用的单值标记，而在MMT模型中，它只是多标记中的一种标记，记为CAT。

CAT可取如下的值：名词、处所词、方位词、区别词、数词、量词、体词性代词、谓词性代词、动词、形容词、副词、介词、连词、助词、语气词、拟声词、感叹词。

为便于计算机处理，我们把标点符号与公式也各算为一个词类，这样一来，汉语共有20个词类，即特征CAT可取20个值。

每个特征值还可以再取子值，即进行进一步的分类。例如，汉语的形容词可以再分为状态形容词和性质形容词两个次类。也就是说，形容词这个值还可再取状态形容词和性质形容词两个子值。特征的值及其子值，可以看成是次一级的“特征/值”偶对，也就是可以把值看成次一级“特征/值”偶对中的特征，把该值的子值看成次一级“特征/值”偶对中的值。这意味着当存在子值时，在“特征/值”偶对中的“值”本身，也可以是一个次一级的“特征/值”偶对。

2.词组类型特征和它的值：

词组类型特征是描述汉语的另一个特征，记为K。在短语结构语法中，它也是常用的单标记，但在同一个结点上，它不能与词类标记共存。在MMT模型中，它只是多标记中的一种标记。

K的值可取：动词词组、名词词组、形容词词组、数量词组，共4个。

我们把传统语法中的介词词组并入名词词组，因为从自然语言计算机处理的角度看来，介词词组中的介词，实际上只是它后

面的名词词组的功能的一种标态，并入名词词组处理更为方便。

3.单词的固有语义特征和它的值：

单词的固有语义特征，就是单词的语义类别，它表示的是孤立的单词的语义，而不是单词与单词之间的语义关系。单词的固有语义特征，记为SEM。

SEM可取如下的值和子值：

物象：其子值为生物、无生物、机关组织、类别名称。

物资：其子值为设备、产品、原材料。

现象：其子值为自然现象、人工现象、社会现象、力能现象。

时空：其子值为时间、空间。

测度：其子值为数量、单位、标准。

抽象：其子值为学问、概念、符号。

属性：其子值为性质、形状、关系、结构。

行动：其子值为行为、动作、操作。

这些固有语义特征都应标注在机器词典中孤立的单词上面，成为单词本身固有的语义属性。

4.单词的固有语法特征和它的值：

孤立的单词也具有语法特征。例如，不同的名词要求不同的量词，因此，带量词特征，就是名词的固有语法特征；不同的动词及物性不同，因此，及物性就是动词的固有语法特征；不同的动词的“价”(valence) 也不尽相同，因此，“价”就是动词的另一个固有语法特征，“价”反映了动词对其前后词语的要求，但它是动词本身的属性，因此，我们把它看成是动词的固有语法特征。

单词的固有语法特征记为GRM。

这样的语法特征的值也可以具有子值，这时，我们可以把值和它的子值作为“特征/值”偶对来处理。例如，动词的固有语法特征的及物性这个值具有两个子值：“及物”和“不及物”，我们可把及物性看成特征，把及物和不及物这两个子值看成它的这个特

征的值。前面我们用过的TRANS＝TV和TRANS＝IV等表示法，正是这样来处理的。

“价”也可以取子值：一价、二价、三价。一价动词只能有一个主语，如“咳嗽”；二价动词可以有一个主语和一个宾语，如“写”；三价动词可以有一个主语、一个直接宾语、一个间接宾语。如“给”。

5.句法功能特征：

由于现代汉语中的词组类型和句法功能之间没有明确的一一对应关系，它们之间的关系极为错综复杂，因此，在汉语句子的自动分析中，必须注意句法功能特征。这些特征都是在句子的自动分析中产生的，而不是单词或词组本身固有的，它们不能直接记在机器词典中。汉语中句子组成成分的句法功能特征记为SF。

SF可取如下的值：主语、谓语、定语、状语、补语、述语、中心语。

SF的值可以有子值。例如，宾语这个值可以有直接宾语和间接宾语两个子值。

6.语义关系特征：

语义关系特征也不是单词本身固有的，而是在计算机自动进行句法语义分析的过程中通过运算得出的。孤立的单词谈不上语义关系，只有两个或两个以上的单词或词组才会产生语义关系。语义关系特征记为SM。

SM可取如下的值：施事、受事、与事、关涉、时刻、时段、时间起点、时间终点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、工具、方式、目的、条件、作用、内容、范围、论题、比较、伴随、判断、陈述、附加等。

SM的各个值还可以分得更细，这样每个值还可以再取子值。

7.逻辑关系特征：

如果把汉语的句子看成一个逻辑命题，那么，在逻辑命题的谓词与它的各个主目语（arguments）之间还存在着逻辑关系。由

于逻辑命题的各个主目语在句子中是由单词或词组来充当的，因而在句子中，单词与单词或者词组与词组之间还存在着逻辑关系。这种关系就是乔姆斯基在转换生成语法标准理论中所说的“题元关系”(θ relation)。逻辑关系用LR表示。

LR的值如下：

主目语0：它是句子的深层主语

主目语1：它是句子的深层直接宾语

主目语2：它是句子的深层间接宾语。

逻辑关系特征的值一般没有子值。

每一个主目语均起一个题元作用，而且只能起一个题元作用；每个题元作用均由一个主目语来充当，而且只能由一个主目语来充当。因此，可以根据主目语的情况来检验所处理的句子在逻辑关系的分析上是否正确，并且揭示出整个句子的逻辑结构。

上面列出的汉语的“特征/值”系统，还不十分完善，有待在实践中进一步补充。

用这样的“特征/值”系统，我们把树形图中的一个单元性的结点分解为非单元性的特征的组合，使单标记的树形图改造成为多标记的树形图，大大地拓广了树形图表达语言信息的能力。这是语言符号的非单元性这一原理在汉语句子自动分析中的体现。

在上面所列举的各类特征中，词类特征、单词的固有语法特征、单词的固有语义特征都是可以在词典中独立地给出来的，它们是单词本身固有的特征，我们把它们叫做静态特征(static features)。而词组类型特征、句法功能特征、语义关系特征、逻辑关系特征并不能表示单词本身的固有特征，它们是单词与单词之间发生联系时才产生出来的特征，我们把它们叫做动态特征(dynamic features)。

在自动句法语义分析中，静态特征是计算机进行运算的基础，计算机依赖于这些预先在词典中给出的静态特征，通过有穷步运算，逐步算出各种动态特征，从而逐步弄清楚汉语句子中各个语

言成分之间的关系，达到自动句法语义分析的目的。

在各种动态特征中，词组类型特征是最容易运算求出的。一般根据树形图中某个结点的直接后裔的词类特征、单词的固有语法特征及单词的固有语义特征等信息，就不难推算出该结点的词组类型特征。句法功能特征则要通过更广泛的上下文信息才能推算求出，而语义关系特征及逻辑关系特征则是最难求出的，往往不是一步求出，而是要通过许多步的演绎和推理，才有可能推算出来。因此，如何根据各种静态特征推算出动态特征，便是汉语自动分析的关键所在。汉语语法和语义的研究应该为这方面的工作提供出有效的规则，在这个领域中，非常需要语言学家、数学家和计算机专家的通力协作。

一般地说，汉语句子的自动分析，应当包含如下的步骤：

1.对输入的汉语句子进行自动切分，确定单词与单词之间的界限。

2.在词典中查出句子中各个单词的静态特征。

3.根据语法规则和语义规则，检查这些静态特征的相容性，把静态特征相容的单词结合成词组。

4.根据语法规则和语义规则，由静态特征和词组类型特征出发，计算出句法功能特征，并进一步计算出语义关系特征和逻辑关系特征。

在检查静态特征的相容性以及由静态特征计算动态特征时，如果两个特征不相容，则不能进行运算，运算失败，如果两个特征相容，则根据有关的语法和语义规则进行运算。由于在特征不相冲突时就可以对特征进行运算，由运算而得出的特征信息必然不断增多，句子各个组成成分所包含的特征越来越丰富，最后求出的各种特征就能比较全面地反映汉语句子的性质。

汉语的自动生成过程与此相反。在从外语到汉语的机器翻译中，一般是根据外语分析得到的有关句法功能、语义关系、逻辑关系的特征，并根据外汉双语言机器词典中提供的有关汉语单词

的静态特征，进行汉语词序的调整及必要的词形变化（如动词和形容词重叠式的变化），最后产生出合格的汉语句子来。

语言符号的这种非单元性，也就是语言符号特征的复杂性。索绪尔在1916年出版的《普通语言学教程》中早就指出："语言可以说是一种只有复杂项的代数。"①他举出德语中名词数的变化Nacht（夜，单数）：Nächte（夜，复数）来说明这个论点。他认为，Nacht:Nächte 这个语法事实可以用a/b这一符号来代表，但是，其中的a、b都不是简单项而是复杂项，它分别从属于一定的系统之下。Nacht有名词、阴性、单数、主格等特征，它的主要元音为a，Nächte有名词、阴性、复数、主格等特征，它的主要元音为ä，结尾加了e，ch的读音从/x/变为/ç/。这样，就可以形成许多对立，所以叫做复杂项。每个符号孤立地看，可以认为是简单项，但是从整体来看，则都是复杂项。索绪尔指出："语言的实际情况使我们无论从哪一方面去进行研究，都找不到简单的东西；随时随地都是这种相互制约的各项要素的复杂平衡。"②可见，索绪尔早就提出了要用"复杂项"来描述语言的观点，他所说的"复杂项"，就是我们现在所说的"多标记"，它们都体现了语言符号的非单元性。索绪尔真不愧是一位慧眼独具的学者，可惜他的这一卓越思想并没有得到后世语言学家的重视。号称继承了索绪尔结构主义语言学思想的美国描写语言学派，在他们提出的"直接成分分析法"中，只采用简单标记来描述句子，而在乔姆斯基的短语结构语法中，则更是明确地用"单标记"来描述句子。现在，当我们用短语结构语法对自然语言进行计算机处理遇到重重困难而感到山穷水尽的时候，我们回过头来重温索绪尔关于"复杂项"的思想，不得不由衷地佩服这位学术前辈的远见卓识。事实上，我们中国的语言学者为了解决在用短语结构语法来描述汉语中碰

---

① 索绪尔，《普通语言学教程》，中译本，商务印书馆，1980年，第169页。

② 索绪尔，《普通语言学教程》，中译本，商务印书馆，1980年，第169页。

到的种种问题，正是从索绪尔关于"复杂项"的思想中得到启示，才提出了"多标记"和"多值标记函数"的概念。由此可以看出语言学的基础理论对于自然语言计算机处理研究实践的指导意义。

第2节 复杂特征的运算

就在中国语言学者提出MMT模型的同时，国外一些计算语言学家也看到了乔姆斯基短语结构语法的局限性，纷纷提出各种手段来限制短语结构语法的过强的生成能力，来提高短语结构语法的有限的分析能力。这些手段中，最为有效的就是"复杂特征"(complex features)和"合一"(unification)。

80年代以来，首先从伍兹(W. Woods)的"扩充转移网络"(Angmented Transition Network，简称ATN)开始①，在布列斯南(J. Bresnan)关于面向词汇的非转换语言学思想的激励之下，卡普兰(R. Kaplan)和布列斯南一起，于1983年提出了"词汇功能语法"(Lexical-functional Grammar，简称LFG)②；马丁·凯依(Martin Kay)于1983年提出了"合一语法"(Unification Grammar，简称LFG)，于1985年提出了"功能合一语法"(Functional Unificational Grammar，简称FUG)③。这些语法都采用了"复杂特征结构"，而"合一"就是对复杂特征进行运算

① W. Wood, *Tansition Network grammar for natural language analysis*, «Communication of the ACM», 13(10), 1970

② R. Kaplan, J. Bresnan, *Lexical-functional grammar. A formal system for grammatical representation*, in «The Mental Represen tation of Grammatical Relations», 1983.

③ M. Kay, *Parsing in functional unification grammar*, in «Natural Language Parsing, Psychological, Computational and Theoretical Perspectives»,1985.

的方法。

柯尔迈洛埃（A. Colmerauer）于1970年独立地研制了Q-系统（Q-system），又于1978年提出了"变形语法"（Metamorphosis Grammar），把它们作为自然语言处理的工具。在逻辑程序设计方面，佩瑞拉（P. Pereira）和瓦楞（D. Warren）于1980年提出了"定子句语法"（Definite Clause Grammar）简称DCG），这种语法是在柯尔迈洛埃早期形式语法的研究以及程序设计语言Prolog的工作的基础上研制而成的。在独立的逻辑程序设计工作中，这种"定子句语法"已成为许多立足于"复杂特征"和"合一"运算的形式化方法的基础，例如，"移位"（extraposition）、"槽"（slot）和"间隔语法"（Gapping Grammar）等等。这些工作也是离不开"复杂特征"的运算的①。

盖兹达(G. Gazdar)、克莱因（E. Klein）、沙格（I. Sag）和普鲁姆(G. Pullum)等人于1985年提出了"广义短语结构语法"(Generalized phrase Structure Grammar，简称GPSG)②，这种语法以短语结构语法作为基础，采用"特征/值"系统来描述句子，在这种"特征/值"系统中，既包括简单特征，也包括复杂特征，这就在很大程度上，限制了短语结构语法过强的生成能力。在他们最近的研究工作中，也引进了"合一"来进行复杂特征的运算。珀拉德（C. Pollard)于1984年在他的博士论文中，提出了"中心词语法"（Head Grammar)③，其理论基础之一就是"广义短语结构语法"中的"特征/值"系统，1985年，珀拉德和他的同事们又

① F. Pereira, P. Warren, *Definite Clause grammar for language analysis -- A survey of the formalism and a comparison with augmented transition networks,* «*Artificial Intelligence*», 1980.

② G. Gazdar, E. Klein, G. Pullum, I. Sag, *Generalized phrase Structure Grammar*, 1985.

③ C. Pollard, *Generalized phrase structure grammar, head grammar, and natural languages, Doctoral dissertation,* 1984.

提出了"中心词驱动的短语结构语法"(Head-driven Phrase Structure Grammar，简称HPSG)①，这种语法是"广义短语结构语法"和"中心词语法"的进一步发展，也采用了"复杂特征"和"合一"运算。

纵观计算语言学的发展历史可以看出，我国学者在1981年提出的MMT模型，是世界各国学者对传统的短语结构语法进行改进的一个重要方面和不可分割的组成部分。"多标记"的概念也就是"复杂特征"的概念，它是80年代计算语言学形式化方法的一个有力的工具。80年代以来的计算语言学，在关键性的地方都使用了基于"复杂特征"的"合一"运算方法，可以说，"复杂特征"的概念，是当代计算语言学的一个关键性概念，它反映了计算机时代人们对语言符号的非单元性的认识进一步深化了。

复杂特征的运算要采用数理逻辑中"合一"运算的方法，语言符号的非单元性使与数理逻辑发生了联系。

"合一"这个术语最初是在数理逻辑中的一阶谓词演算开始使用的。马丁·凯依的"功能合一语法"，在名称上冠以了"合一"的字眼儿，因此，我们通过对"功能合一语法"的介绍，便不难理解"合一"运算在语言学中的实际运用情况。

马丁·凯依于1985年在"功能合一语法"这一新的语法理论中，提出了"复杂特征集"(complex feature set)的概念。他认为，自然语言是一个效率极高同时又能够精确地表达各种意念的信息系统，仅只用乔姆斯基的短语结构语法中的单标记的句法范畴不可能充分地描述自然语言的句子结构，而必须使用复杂特征集来描述。②

① D. Proudian, C. Pollard, *Parsing head-driven Phrase Structure grammar*, in «*Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*», 1985.

② 黄昌宁，《机器翻译与新的语法理论》，(《中国计算机用户》)，1989年，第9期。

复杂特征集用功能描述（Functional Description，简称FD）来表示。功能描述FD由一组描述元（descriptors）组成，而每一个描述元则是一个成分集（constituent set）、一个模式(Pattern)或一个带值的属性（attribute），其中最主要的是“属性/值”偶对。在功能描述FD中，描述元的值可以是原子，也可以是另一个功能描述FD。所以，功能描述是递归地定义的。

下面给出表示复杂特征集的功能描述的数学定义：

α为一个功能描述FD，当且仅当α可表示为

$$\begin{bmatrix} f_1 = v_1 \\ f_2 = v_2 \\ \vdots \\ f_n = v_n \end{bmatrix} \quad n \geqslant 1$$

其中，f_i表示特征名，v_i表示特征值，而且，满足如下条件：

① 特征名f_i为原子，特征值v_i或为原子，或为另一个功能描述FD；

② $\alpha\langle f_i \rangle = v_i$，（$i = 1, \cdots, n$）

读作：集α中，特征f_i的值等于v_i。

采用这样的功能描述，就可以表示复杂特征集。

组成功能描述FD的一组描述元都写在一个方括号里，书写的顺序无关紧要。在一个“属性/值”偶对中，属性是一个符号，如NUMBER(数)、SUBJ（主语）、OBJE（宾语）、MODF（修饰语）、HEAD(中心语）等，它的值或者是一个符号，或者是另外一个功能描述FD，属性和它的值之间用等号来连接，因此，$a = b$表示属性a的值是b。

例如，英语句子We helped her（我们帮助过她)可以用（1）所示的功能描述FD来表示：

(FD1):

$$
\left[\begin{array}{l}
K = S \\
SUBJ = \left[\begin{array}{l} CAT = PRON \\ CASE = NOM \\ NUMBER = PLUR \\ PERSON = FIRST \end{array}\right] \\
OBJE = \left[\begin{array}{l} CAT = PRON \\ GENDER = FEM \\ CASE = ACC \\ NUMBER = SING \\ PERSON = THIRD \end{array}\right] \\
PRED = \left[\begin{array}{l} CAT = VERB \\ LEX = \text{‘help’} \end{array}\right] \\
TENSE = PAST \\
VOICE = ACTIVE
\end{array}\right]
$$

这个功能描述表示："We helped her"是个句子 (K=S)，在这个句子中，主语"we"是代词、主格、复数、第一人称，宾语"her"是代词、阴性、宾格、单数、第三人称，谓语"helped"是动词，具体的词是"help"，整个句子的时态是过去时，语态是主动态。这些功能描述也就是这个句子的复杂特征集。

在一个功能描述FD中，每一个"属性/值"偶对都是该FD所描述对象中的一个特征。如果这个值是一个符号，那么，这个"属性/值"偶对就叫做功能描述FD的一个基本特征。任何功能描述FD都可以用一张由基本特征组成的表来表示。例如，上面的功能描述FD(1)也可以用下面的表FD (2)来描述：

FD(2)　⟨K⟩ = S

⟨SUBJ　CAT⟩ = PRON

⟨SUBJ　CASE⟩ = NOM

⟨SUBJ NUMBER⟩ = PLUR

⟨SUBJ PERSON⟩ = FIRST

⟨OBJE CAT⟩ = PRON

⟨OBJE GENDER⟩ = FEM

⟨OBJE CASE⟩ = ACC

⟨OBJE NUMBER⟩ = SING

⟨OBJE PERSON⟩ = THIRD

⟨PRED CAT⟩ = VERB

⟨PRED LEX⟩ = 'help'

⟨TENSE⟩ = PAST

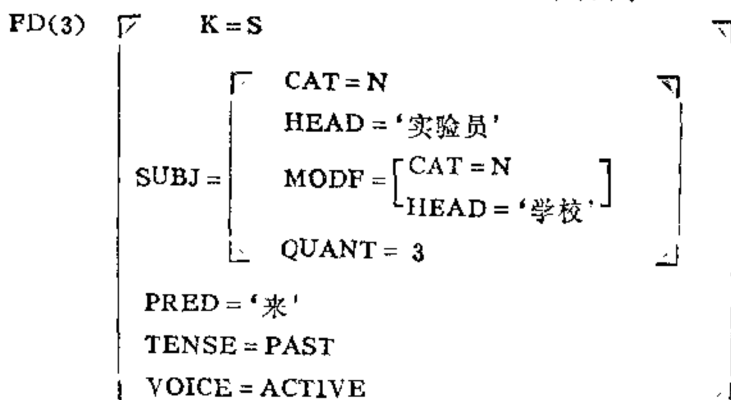
⟨VOICE⟩ = ACTIVE

在这个表FD(2) 中，尖括号⟨ ⟩里的符号构成了一条路径(Path)，功能描述FD中的每一个值， 都可以用一条路径来称呼它。可以看出，FD(2)中表达的特征与FD(1)中表达的特征是相同的，它们是同一个句子中的复杂特征的不同的表达方式。不过，尽管FD(1)和FD(2)都是同一功能描述FD的两种表示，它们还各有不同，FD(1)显示了功能描述的嵌套，因而强调了功能描述的结构特性，FD(2)是一个表，因而强调了功能描述的内部分量特性。这两种表示方法都有意模糊了特征和结构之间的通常区别，使得功能合一语法具有更大的灵活性。我们在MMT模型中对复杂特征的表示方法，与这里的FD(2)比较接近，因为MMT模型对于结构层次的描述，是通过多叉树来表示的，所以，在只描述句子的代数值的复杂特征中，就没有必要再强调结构特性的描述了。

把功能描述看作是非结构性的特征集，就有可能用集合论的标准运算来处理它们。但是，功能描述又不能完全服从集合论的运算。集合论运算一般并不考虑运算对象的相容性，而功能描述则必须考虑运算对象的相容性。如果有两个功能描述中都包含一个共同的属性，而这个共同属性在这两个功能描述中的值（可以是符号，也可以是另外的功能描述FD）不相同，那么，这两个功

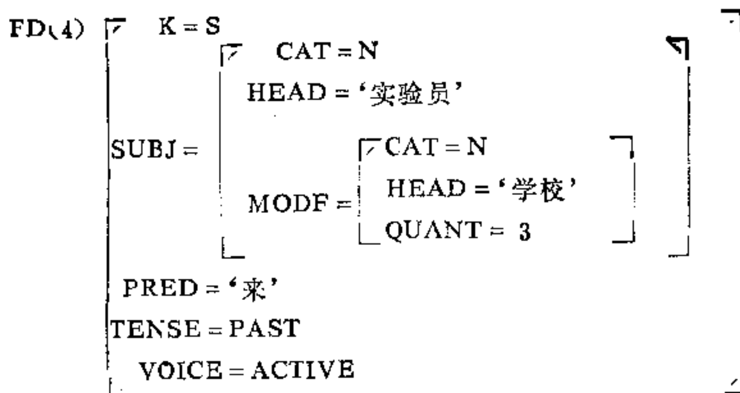
能描述就是不相容的。例如，如果功能描述 F1 中含有基本特征⟨A⟩ = x，功能描述F2中含有基本特征⟨A⟩ = y那么，除非$x = y$，否则，F1与F2不相容。如果两个功能描述不相容，那么，在进行集合论中的“并”运算时，运算的结果就不会是一个合格的功能描述。例如，假定功能描述F1所描述的句子中含有一个单数主语，而功能描述F2所描述的句子中含有一个复数主语，那么，如果S1和S2是它们相应的基本特征集，它们的并集 S1∪S2就是不合格的，因为在这个并集中，⟨SUBJ NUMBER⟩ = SING 和 ⟨SUBJ NUMBER⟩ = PLUR不相容。

对于语法上有歧义的句子或词组，需要两个或两个以上的不相容的功能描述来表示。例如，“三个学校的实验员来了”这个句子是有歧义的，它有两个不同的意思。一个意思可用功能描述FD(3)来表示，另一个意思可用功能描述FD(4)来表示：

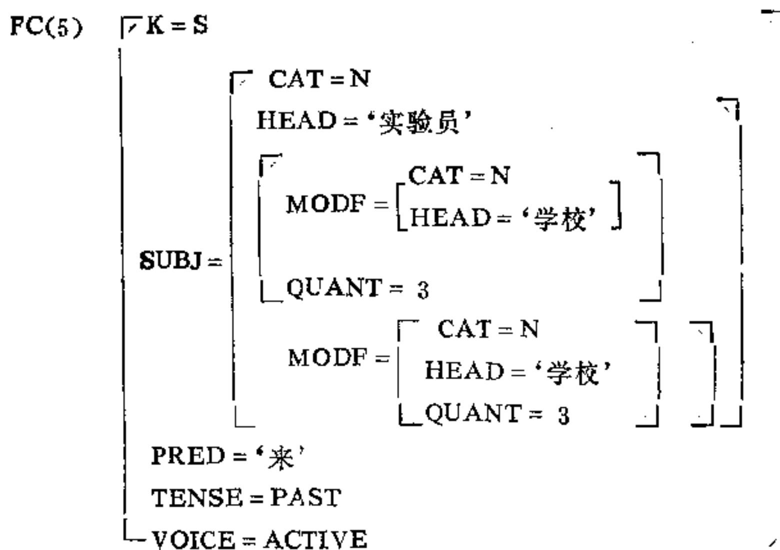


可以看出，在FD(3) 中，句子的意思是只来了3个实验员，而这3个实验员是学校的实验员；在FD(4) 中，句子的意思是来了一些实验员，而这些实验员分属3个不同的学校。

几个不相容的简单的功能描述FD：F_1，…，F_k，可以合并成一个单独的复杂的功能描述FD：{F_1…，F_k}，复杂的功能描述表示分量的对象集的并，其中的不相容部分，应用花括号括起来



下面是把FD(3)和FD(4)合并而成的复杂的功能描述FD(5),它描述了FD(3) 和FD(4) 所分别表示的两种结构关系：



FD(5)中的花括号表示不相容的功能描述或子功能描述之间的析取关系。用这种复杂功能描述的紧凑形式，可以描述大量的互不相容的对象。一般地说，功能合一语法中的语法规则可以用

一个统一的功能描述FD(6)表示如下：

$$
\text{FD}(6)\quad \left[\left\{\begin{array}{c}
\left[\begin{array}{c}\text{CAT}=C_1\\ \vdots\end{array}\right]\\
\left[\begin{array}{c}\text{CAT}=C_2\\ \vdots\end{array}\right]\\
\vdots\\
\left[\begin{array}{c}\text{CAT}=C_n\\ \vdots\end{array}\right]
\end{array}\right\}\right]
$$

对于采用这种复杂特征集来描述的系统来说，其描述的详尽程度是没有限制的。一个描述中所包含的特征越多，它对所描述的对象的限定也就越具体；如果从一个描述中撤消某些特征，就可能扩大它所描述的对象的覆盖面。因此，灵活地控制特征的数量，认真地选择特征的内容，才可以用复杂特征集进行恰当的描述。

在机器翻译的机器词典中，对于每一个单词的定义，不仅仅给出其词类，而且，还应该标出这个词的静态的词法特征、句法特征和语义特征，这就是在词这一级采用复杂特征集。随着自动句法分析的推进和自动语义分析的进行，句子中的每个单词除了被标注上来自词典中的这些静态特征之外，在表示句子层次结构的树形图的每个结点上，计算机还会运算出一些动态特征，它们大大地充实了来自词典中的静态特征的内容，这些动态特征当然也要以复杂特征集的形式来标注，这就是在句法分析和语义分析这一级采用复杂特征集。复杂特征集中的各种复杂特征，可以在短语归并的过程中从中心词的复杂特征标记中继承过来，也可以根据句法语义规则动态地通过计算机计算出来。在原语自动分析中采用这样的复杂特征集，有效地解决了兼类词和歧义结构的判定等困难问题，并且把句法分析和语义分析通过复杂特征集这种手段有机地结合起来，从而提高了原语句法语义分析的效率。

自然语言自动句法语义分析中的复杂特征，比之于雅可布逊提出的音位学中十二对区别特征，要丰富得多，它们不仅是二元对立的，而且还是多元对立的，不仅具有线性的结构，而且还具有嵌套的、递归的结构，所以，对于复杂特征集，就不能采用一般的“并”运算方法来进行运算，而要采用新的方法。

功能合一语法是采用“合一”这种独特的运算方式来对复杂特征集进行运算的。

“合一”是数理逻辑中的一阶谓词演算使用的一种运算方法。

寻找某种项对变量的置换，从而使表达式一致的过程叫做合一。如果存在一个置换S，把它作用到表达式集$\{E_i\}$中的每一个元素上，使得$E_{1s}=E_{2s}=\cdots=E_{ns}$，那么，就说表达式集$\{E_i\}$是可合一的，S就叫做$\{E_i\}$的合一者 (unifier)，因为它的作用是使该集合简化为一致的形式。

例如，有两个逻辑项

A：$f(x, y)$

和　B：$f(g(y, a), h(a))$，

如果用逻辑项

C：$x=g(h(a), a)$

和　D：$y=h(a)$

来置换A、B中的变量x和y，则置换之后，A和B均成为$f(g(h(a), a)h(a))$，从而使A和B都具有一致的形式，这个结果叫做A和B的合一，C和D叫做A、B的合一者，A、B叫做可合一的逻辑项。

目前，这种合一运算已经被广泛地应用于高阶逻辑、计算复杂性理论、可计算性理论、逻辑程序设计等领域，并进一步被应用到计算语言学、机器翻译、自然语言理解和人工智能等领域。合一运算被如此广泛应用的原因之一是逻辑程序设计语言PROLOG的普及，因为PROLOG语言在霍恩子句 (Horn clause)的归结过程中所依据的基本运算之一就是合一运算。

在功能合一语法中，使用合一运算来把若干个功能描述FD合

并成一个单独的功能描述 FD。具体地说，如果有两个或两个以上简单的功能描述是相容的，便可通过合一运算把它们合并成一个简单的功能描述，使得这个功能描述所描述的对象，正是对面若干个功能描述所共同描述的对象。

这样的合一运算与集合论中的求并运算十分相似，但合一运算与求并运算的不同之处在于，当合一被应用于不相容的项时，合一失败，并产生一个空集。

求并运算所得到的并集是参与运算的各个集合里所有不同元素组成的集合。例如，

$\{A, B\} \cup \{C, B\} = \{A, B, C\}$

在求并运算时，总是把集合中的元素看成是不可分解的原子，即使元素是有序的偶对，如（f_i，v_i）偶对表示特征f_i的值为v_i，求并运算时仍然把它们看成是不可再分解的个体，而不考虑它们的内部结构。假设

$$\alpha = \{(f_1, v_1), \quad (f_2, v_2)\}$$

$$\beta = \{(f_1, v_1')\}$$

即使$v \neq v'$，α与β所表达的信息互相抵触，在进行求并运算之后，其并集仍然为

$\gamma = \alpha \cup \beta = \{(f_1, v_1), (f_1, v_1'), (f_2, v_2)\}$ 在并集中虽然保持了抵触的信息，不过，从信息组合和传递的角度来看，所求得的并集γ是没有意义的。

合一运算必须考虑运算结果的合理性，在合一运算中，当α与β所表达的信息相互抵触时，其合一的结果为空集（记为$\varnothing$），表示合一失败，如果用符号$\overline{\cup}$表示合一，则有

$$\alpha \overline{\cup} \beta = \varnothing$$

可见，合一运算与求并运算得到的结果是很不相同的。

下面我们给出在功能合一语法中合一运算的形式定义：

如果有某种运算$\overline{\cup}$具有如下性质

1. 若a和b均为原子，则$a \overline{\cup} b = a$，当且仅当$a = b$；否则$a \overline{\cup} b = \varnothing$。

2. 若α和β均为复杂特征集，则

①若$\alpha(f) = \nu$，但$\beta(f)$的值未经定义，则$f = \nu$属于$\alpha \overline{\cup} \beta$；

②若$\beta(f) = \nu$，但$\alpha(f)$的值未经定义，则$f = \nu$属于$\alpha \overline{\cup} \beta$；

③若$\alpha(f) = \nu_1$，$\beta(f) = \nu_2$，且ν_1与ν_2不相抵触，则

$f = (\nu_1 \cup \nu_2)$属于$\alpha \overline{\cup} \beta$；否则，$\alpha \overline{\cup} \beta = \varnothing$。

那么，就把这种运算$\overline{\cup}$叫做合一运算。

从合一运算的定义可以看出，集合论中的求并运算是合一运算的一种特殊情况。当合一的对象所含的元素为不可分解的原子时，合一的结果等于并集。当合一的对象是有结构的复杂特征集时，就要检验特征的相容性，只有当特征相容时，相应的复杂特征才能合一。因此，合一运算具有两种作用：

第一，合并原有的特征信息，构造新的特征结构，这与集合论中的求并运算类似；

第二、检查特征的相容性和规则执行的前提条件，如果参与合一的特征相冲突，就立即宣布合一失败。

可见，合一运算提供了一种在合并各方面来的特征信息的同时，检验限制条件的机制。这正是语言符号的非单元性在计算机运算方面所需要的，因此，这种合一运算受到了计算语言学工作者的欢迎。

我们举例说明如何进行合一运算。

例1.

$$\begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{LEX} = \text{'run'} \\ \text{TENSE} = \text{PRES} \end{bmatrix} \overline{\cup} \begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{NUMBER} = \text{SING} \\ \text{PERSON} = \text{THIRD} \end{bmatrix} \rightarrow$$

$$\begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{LEX} = \text{'run'} \\ \text{TENSE} = \text{PRES} \\ \text{NUMBER} = \text{SING} \\ \text{PERSON} = \text{THIRD} \end{bmatrix}$$

由于参与合一运算的两个功能描述中的复杂特征是相容的，因此，合一运算的结果等于这两个功能描述中的复杂特征求并。

例2.

$$\begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{LEX} = \text{'run'} \\ \text{TENSE} = \text{PRES} \end{bmatrix} \bar{\cup} \begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{TENSE} = \text{PAST} \\ \text{PERSON} = \text{THIRD} \end{bmatrix} \rightarrow \text{NIL}$$

由于这两个功能描述中，第一个功能描述中的TENSE = PRES，第二个功能描述中的TENSE = PAST，相互抵触，因而合一运算的结果为NIL，表示合一失败。

例3.

$$\begin{bmatrix} \left\{ \begin{matrix} \text{TENSE} = \text{PRES} \\ \text{FORM} = \text{'is'} \\ \text{TENSE} = \text{PAST} \\ \text{FORM} = \text{'was'} \end{matrix} \right\} \end{bmatrix} \bar{\cup} \begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{TENSE} = \text{PAST} \end{bmatrix} \rightarrow$$

$$\begin{bmatrix} \text{CAT} = \text{VERB} \\ \text{TENSE} = \text{PAST} \\ \text{FORM} = \text{'was'} \end{bmatrix}$$

第一个功能描述是由不相容的两个简单功能描述合并而成的复杂功能描述，它与第二个功能描述进行合一运算时，取相容的特征作为合一运算的结果。由于第一个复杂功能描述中的特征

$$\begin{bmatrix} \text{TENSE} = \text{PRES} \\ \text{FORM} = \text{'is'} \end{bmatrix}.$$

与第二个功能描述中的特征不相容，故被舍去。

一般地说，两个复杂功能描述的合一运算结果仍然还是复杂

功能描述，其中，每一项代表原来的功能描述中的一对相容项。因此，

$$\{a_1, a_2, \cdots, a_n\} \bar{\cup} \{b_1, b_2, \cdots, b_m\}$$

就得到一个形式为 $\{c_1, c_2, \cdots, c_k\}$ 的功能描述，其中每一个c_h $(1 \leqslant h \leqslant k)$ 都是一对相容项的合一运算结果 $a_i = b_j$ $(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m)$。

由此可见，合一运算应该具有如下的性质：

1. 合一运算可以对信息进行相加：

例如，

$$[\mathrm{CAT=N}] \bar{\cup} [\mathrm{AGREEMENT=[NUMBER=SING]}]$$

$$\rightarrow \begin{bmatrix} \mathrm{CAT=N} \\ \mathrm{AGREEMENT=[NUMBER=SING]} \end{bmatrix}$$

2. 合一运算是幂等的：

例如，

$$[\mathrm{CAT=N}] \bar{\cup} \begin{bmatrix} \mathrm{CAT=N} \\ \mathrm{AGREEMENT=[NUMBER=SING]} \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \mathrm{CAT=N} \\ \mathrm{AGREEMENT=[NUMBER=SING]} \end{bmatrix}$$

前一个复杂特征集中的CAT = N被吸收到后一个复杂特征集当中去了。

3. 空白项是合一运算的幺元：

$$[\quad] \bar{\cup} \begin{bmatrix} \mathrm{CAT=N} \\ \mathrm{AGREEMENT=[NUMBER=SING]} \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \mathrm{CAT=N} \\ \mathrm{AGREEMENT=[NUMBER=SING]} \end{bmatrix}$$

空白项与复杂特征集进行合一运算，则该空白项被复杂特征集吸收。

4. 当特征值相容时，相同的特征可以合一：

例如，

$$\begin{bmatrix} \text{AGREEMENT} = [\text{NUMBER} = \text{SING}] \\ \text{SUBJ} = [\text{AGREEMENT} = [\text{NUMBER} = \text{SING}]] \end{bmatrix}$$

$$\bar{\cup}[\text{SUBJ} = [\text{AGREEMENT} = [\text{PERSON} = \text{THIRD}]]]$$

$$\rightarrow \begin{bmatrix} \text{AGREEMENT} = [\text{NUMBER} = \text{SING}] \\ \text{SUBJ} = \left[\text{AGREEMENT} = \begin{bmatrix} \text{NUMBER} = \text{SING} \\ \text{PERSON} = \text{THIRD} \end{bmatrix}\right] \end{bmatrix}$$

由于在前后复杂特征集中，特征SUBJ和特征 AGREEMENT 的特征值 NUMBER = SING和 PERSON = THIRD 是相容的，所以，合一后形成特征

$$\left[\text{SUBJ} = \left[\text{AGREEMENT} = \begin{bmatrix} \text{NUMBER} = \text{SING} \\ \text{PERSON} = \text{THIRD} \end{bmatrix}\right]\right]$$

如果把自然语言看作是一个传递和负载信息的系统，并且承认自然语言中的句法成分和语义成分都可由较小的成分合成较大的成分，那么，采用合一作为句法和语义分析的基本运算便是非常理想的了。这是因为，

第一、一个语言单位（如句子或词组等）所负载的信息可以分布在各个成分之中，每个成分所负载的可以只是部分的信息。

第二、通过合一运算，在小成分组合成大成分的过程中，小成分所负载的信息也同时被传递或累加为大成分所负载的信息，在合一运算的过程中，信息只会逐渐增加而不会减少。

第三、由于句法分析和语义分析都以合一运算作为基本运算，不仅句子的合法性可以通过语义手段来判断，而且，还可以把句子的句法结构和语义表示用合一运算这种方式更加自然地衔接起来。

第四、不同的功能描述的合一运算结果，同这个运算所进行的先后次序无关，不论合一从哪个方向开始，也不论是先合一还是后合一，合一的结果都是相同的。合一运算的这种无序性非常

便于进行并行处理，而且还使我们有可能自由地选择分析算法和自然语言描述的语法理论。

在复杂特征集与合一运算的基础上，马丁·凯依提出了功能合一语法。

功能合一语法的最大特点就是在词条定义、句法规则、语义规则和句子的描述中，全面地、系统地使用复杂特征集。

1.词条定义的描述：

例如，英语的saw有三个义项，在词条saw中，可给出三条定义，每一条定义的形式都是复杂特征集的功能描述。见FD(7)、FD(8)和FD(9)。

FD(7)：

CAT = V
TENSE = PAST
TRANSITIVITY = MENTAL - PROCESS
ROOT = 'see'
LEX = 'saw'

FD(7)表示saw是动词see的过去时形式，它的含义是"看见"。

FD(8)：

$$\begin{bmatrix} \text{CAT} = \text{N} \\ \text{NUMBER} = \text{SING} \\ \text{LEX} = \text{'saw'} \end{bmatrix}$$

FD(8)表示saw是名词，它的含义是"锯子"。

FD(9)：

$$\begin{bmatrix} \text{CAT} = \text{V} \\ \text{TENSE} = \text{INFINITIVE} \\ \text{TRANSITIVITY} = \text{MATERIAL - PROCESS} \\ \text{ROOT} = \text{'saw'} \\ \text{LEX} = \text{'saw'} \end{bmatrix}$$

FD(9)表示saw是动词saw的不定式形式，它的含义是“锯”。

2.句法规则的描述：

例如，FD(10)和FD(11)分别是主动态和被动态的规则：

FD(10)：

```
[ K = S
  PATTERNS = (…PREDICATE DIRECT－OBJECT…)
  SUBJ = ACTOR = [CAT = N]
                [ CAT = V
  PREDICATOR =    TRANSITIVITY = MATERIAL
                                 －PROCESS
                  VOICE = ACTIVE ]
  VOICE = ACTIVE ]
```

FD(11)：

```
[ K = S
  PATTERNS = (…PREDICATOR…BY…ADJUNCT…)

  SUBJ = AFFECTED = [CAT = N]
                [ CAT = V
  PREDICATOR =    TRANSITIVITY = MATERIAL
                                 －PROCESS
                  VOICE = PASSIVE ]
                   [ K = PP
  BY－ADJUNCT =      PREP = [ CAT = PREPOSITION
                              LEX = 'by' ]
                     OBJE = <AGENT> ]
  VOICE = PASSIVE ]
```

其中，ACTOR表示施事，AFFECTED表示受事，其它符号的含义从相应的英文词的词义不难体会出来。

这两条规则的调用条件是：

1. 句法成分的K = S，即它们应是句子(sentence)；

2. 调语动词表示一个“物质过程”，即

TRANSITIVITY = MATERIAL – PROCESS

特征 PATTERNS 的值是有序的，它规定了主动态和被动态句型中语言成分的基本顺序，主动态中的 PATTERNS 是（…PREDICATOR DIRECT – OBJECJ…），被动态中的 PATTERNS 是（…PREDICATOR…BY – ADJUNCT…）。这样，根据特征PATTERNS的值，就可以安排和调整有关语言成分的位置。

3. 句子结构的描述：

例如，英语句子She smashed a brick（她砸碎了一块砖）的句子结构可用FD(12)来描述：

FD(12)：

```
┌ K = S
│ PATTERNS = (SUBJ PREDICATOR DIRECT
│                               – OBJECT)
│ TENSE = PAST
│ VOICE = ACTIVE
│                 ┌ K = NP
│                 │ PATTERNS = (HEAD)
│                 │         ┌ CAT = PRON
│                 │         │ GENDER = FEM
│                 │ HEAD =  │ CASE = NOM
│ SUBJ = ACTOR =  │         │ NUMBER = SING
│                 │         │ PERSON = THIRD
│                 │         └ LEX = 'she'
│                 │ NUMBER = SING
│                 │ DEFINITENESS = DEFINITE
│                 └ PERSON = THIRD
```

```
            ┌CAT = V                          ┐
            │TRANSITIVITY = MATERIAL           │
PREDICATOR =│                  - PROCESS       │
            │VOICE = ACTIVE                    │
            └LEX = 'smashed'                   ┘
             ┌K = NP                                    ┐
             │PATTERNS = (DETERMINER                    │
             │                  HEAD                    │
             │           ┌CAT = ARIICLE     ┐           │
             │           │NUMBER = SING     │           │
             │DETER -    │DEFINITENESS      │           │
DIRECT - OBJECT  MINER = │ = INDEFINITE     │           │
  = AFFECTED =│          └LEX = 'a'         ┘           │
             │         ┌CAT = N             ┐           │
             │HEAD =   │NUMBER = SING       │           │
             │         └LEX = 'brick'       ┘           │
             │NUMBER = SING                             │
             │DEFINITENESS = INDEFINITE                 │
             └PERSON = THIRD                            ┘
```

在这个功能描述中，不仅包括了对单词、词组和句子等各语言成分的特征和功能的描述，而且，还说明了中心动词 smashed 的施事（actor）、受事（affected）等语义关系方面的内容。

由于语言符号具有非单元性，而复杂特征集和合一运算的方法，特别适合于描述语言符号的这种非单元性，因而这样的方法已成为了现代计算语言学的主流。除了功能合一语法之外，现代计算语言学的主要流派，如广义短语结构语法、词汇功能语法、中心词驱动的短语结构语法、定子句语法等，都采用了这样的方法。

广义短语结构语法是以上下文无关的短语结构语法作为基础

的，它的信息表达方式就是一个限制的“特征/值”系统，所有的句法特征都是由‹特征，特征值›这样的偶对构成的。特征有两种性质：一是它能有什么样的值；二是它与其它特征在分布上显现什么样的规律性。

一些特征具有终极值。例如，在英语中有如下特征及其终极的特征值：

特征	特征值
PERSON（人称）	{1, 2, 3}
PLUR（复数）	{+, −}
CASE（格）	{NOM, ACC}
VFORM（动词形式）	{FIN,INF,BAS,PAS,…}
PFORM（介词形式）	{to, by, for, …}

其中，NOM表示主格，ACC表示宾格，FIN表示定式动词，INF表示不定式动词，BAS表示原形动词，PAS表示被动式动词。

另一些特征以某个句法范畴为其值，因此它的特征值就是这个句法范畴所具有的特征及这个句法范畴的特征值。例如，特征AGREEMENT就是以句法范畴NP为其值，如果句法范畴NP含有如下特征：

{‹N, +›, ‹V, −›, ‹PERSON, 3›, ‹PLUR, −›},

那么，表示一致关系的特征AGREEMENT的值就是：

{‹AGREEMENT, {‹N, +›, ‹V, −›‹PERSON, 3›,
‹PLUR, −›}›}

由于采用了这样的复杂特征，就能够充分地表达句子中所包含的各种信息，大大提高了乔姆斯基的短语结构语法的描述能力。乔姆斯基曾宣称短语结构语法不适合于以数学的语言来描述自然语言的句子结构，而盖兹达等人则指出，乔姆斯基之所以得出这样的结论，是因为他对短语结构语法的形式化作了不必要的限制，规定只用简单标记，排除了对复杂特征的使用。盖兹达认为，如

果采用复杂特征对原有的短语结构语法进行改造，把短语结构语法发展成广义短语结构语法，而不是象乔姆斯基那样摆脱短语结构语法去搞生成转换语法，那么，这种采用复杂特征的广义短语结构语法将具有生成转换语法的普遍性和生成性，同时还可保留短语结构语法的各种优点。

词汇功能语法把句法结构分为成分结构和功能结构两层。成分结构是语言的外部结构，它表示单词的形式、形态、单词之间的组成方式、短语之间的组成方式等。功能结构是语言的内部结构，它表示谓词的各个主目语（论元）的句法功能、代词的照应关系等等，它本身可以表示为一个属性值矩阵，如下表所示：

属性	值
A	a
B	b
C	c

在这个属性值矩阵中，第一列A，B，C等表示属性，第二列a，b，c等表示相应属性所取的值。这种属性值矩阵实际上就是一个递归的“特征/值”系统。

除此之外，词汇功能语法还带有特殊类型的特征和信息，并且在词汇一级也采用了复杂特征集。词汇功能语法的功能等式实现了复杂特征集在句法结构的各个结点之间的组合和传递。

卡普兰和布列斯南证明了，在词汇功能语法中，由成分结构到功能结构的运算在数学上是有定解的（decidable），而且所有的运算都只需要采用合一来进行。

中心词驱动的短语结构语法通过引入环绕中心词的符号运算，放宽了广义短语结构语法中对上下文无关的特征系统的某些限制，扩充了广义短语结构语法的描述能力，由于整个句子是以中心词为核心而把复杂特征集的信息联系起来的，复杂特征集在这种语法中起着举足轻重的作用。

近年来，逻辑语法有了很大的发展。逻辑语法（logic gram.

mar）是指用谓词逻辑来表达的语法，它是逻辑程序设计和现代语言学相结合的产物。在机器翻译和自然语言理解的研究领域里，经常使用谓词逻辑来描述知识和进行逻辑推理。70年代以来，逻辑以PROLOG语言作为形式被应用于程序设计，谓词逻辑就不再仅仅用于描述知识和逻辑推理的问题，还作为逻辑程序设计的工具来描述解决问题的过程。PROLOG语言使得逻辑和程序设计这两个相距甚远、完全不同的概念协调统一为一个概念——逻辑程序设计。在用PROLOG语言来解决机器翻译和自然语言理解的各种问题的研究过程中，逻辑语法日益成熟起来。

目前主要有四种影响较大的逻辑语法：定子句语法（Definit Clause Grammar，简称DCG），外位语法（e Xtraposition Grammar，简称XG），修饰成分结构语法（Modifier Structure Grammar，简称MSG），约束逻辑语法（Restricting Logic Grammar，简称RLG）。这些语法都在不同程度上突破了短语结构语法只采用简单特征来描述语法的限制。由于篇幅的限制，我们以定子句语法为例来说明这个问题。

瓦楞和佩瑞拉于1980年提出的定子句语法是一种仅仅使用短语结构语法规则的逻辑语法。定子句语法的基本思想是：语法的符号不仅仅是原子符号，还可以是广义的逻辑项。例如，短语结构语法的规则

Sentence ⟹ noun⌒phrase, Verb⌒phrase

表示一个句子由名词短语和动词短语两部分组成，在定子句语法中，同样这个规则可以表示：如果存在一个名词短语和一个动词短语，那么，就存在一个句子的推理过程。短语结构语法的规则与定子句语法的规则在形式上虽然有许多相同之处，但是在本质上却有很大的区别，短语结构语法只是用于描述一种语言，而定子句语法则可用来进行语言的推理。这样，定子句语法便实现了从描述性的形式语法到推理性的逻辑语法的转变，从而使短语结构语法产生了质的飞跃。

在逻辑程序设计中，提出了所谓"霍恩子句"(Horn Clause)。霍恩子句就是一种至多只含有一个正文字的短句，正文字是为原子公式的文字，因此，在霍恩子句中，至多只含有一个为原子公式的文字，这个为原子公式的文字一般写在霍恩子句的左部。霍恩子句的形式为：

$$P: Q_1, Q_2, \cdots, Q_n$$

其中，P是正文字，即原子公式的文字，Q_1，Q_2，…，Q_n都不是正文字。

霍恩子句逻辑意义清晰、形式简明，给程序设计带来很大的方便。从逻辑程序设计的观点来解释，可把霍恩子句看成是左部至多只含有一个谓词的规则。例如，上面的定子句语法规则用霍恩子句可写为：

```
sentence(S0, S):  - noun⌒phrase(S0, S1),
                       verb⌒phrase(S1, S)
```

这里，S0，S1，S为字符串的指针。这个霍恩子句可解释为：如果S0到S1之间是一个名词短语，S1到S之间是一个动词短语，那么，S0和S之间就是一个句子。可见，霍恩子句具体地反映了句子的推理过程。

由于定子句语法中的符号是逻辑项，这就使得定子句语法规则中的非终极符号可以携带有关上下文、转换、结构等多方面的信息，大大地增强了短语结构语法描述自然语言复杂特征的能力。而且，定子句语法规则的右部不仅可以是终极符号和非终极符号，还可以带测试条件的信息，便于描述自然语言的规律。这种带有多方面信息的描述，必须使用复杂特征集和合一运算的方法。定子句语法虽然在形式上使用了短语结构语法，但是，它的描述能力已经相当于乔姆斯基定义的0型文法。所以，我们认为，定子句语法是采用逻辑程序设计的观点以及复杂特征集和合一运算的方法对乔姆斯基短语结构语法的一个重要改进，这是语言符号的非单元性的又一个有力的证明。

第七章

语言符号的模糊性与模糊数学

第1节　语言符号的模糊性

索绪尔没有认识到语言符号具有模糊性。他在《普通语言学教程》一书中说，“从心理方面看，思想离开了词的表达，只是一团没有定形的、模糊不清的浑然之物。哲学家和语言学家常一致承认，没有符号的帮助，我们就没法清楚地、坚实地区分两个观念。思想本身好象一团星云，其中没有必然划定的界限。预先确定的观念是没有的。在语言出现之前，一切都是模糊不清的”①。他又说，“语言对思想所起的独特作用不是为表达观念而创造一种物质的声音手段，而是作为思想和声音的媒介，使它们的结合必然导致各单位之间彼此划清界限。”② 显然易见，索绪尔认为，正是由于语言的作用才使模糊的思想和声音的各单位之间清晰起来，他完全没有认识到语言本身也具有模糊性。

① 索绪尔，《普通语言学教程》，中译本，商务印书馆，1980年，第157页。

② 同①，第157—158页。

但是，早在古希腊时代，语言中的模糊现象就引起了人们的注意。古希腊哲学迈加拉学派的代表人物之一尤布里德斯（Eubulides）就提出了著名的"连锁推理悖论"，这个悖论以多种形式流传下来，其中的一种可叙述如下：

"一粒麦子构不成一堆，对于任何一个数字n来说，如果n粒麦子形不成堆的话，那么，$n+1$粒麦子也形不成堆，因此，任意多的麦粒也形不成堆"。

这个悖论利用了"堆"这个概念的模糊性，因为多少麦粒可以构成一"堆"，是模糊的，而且，n粒麦子与$n+1$粒麦子是否构成一"堆"的界限也是模糊的，所以，人们很容易轻信这个悖论所进行的推理。

从尤布里德斯以后的两千年左右，人们严重地忽视了自然语言词语的模糊性。直到1902年，美国数理逻辑学者皮尔斯（Pearce）又开始研究"模糊"问题，并给模糊下了这样的定义："当事物出现几种可能的状态时，尽管说话者对这些状态进行了仔细的思考，实际上仍不能确定，是把这种状态排除某个命题，还是归属这个命题。这时候，这个命题就是模糊的。上面说的实际上不能确定，我指的并不是由于解释者的无知而不能确定，而是因为说话者的语言的特点就是模糊的"①。1908年，德国学者安东·马尔蒂（Anton Marty）在《普遍语法和语言哲学基础研究》一书中，对语言的模糊性发表过深刻的见解，他指出，"我们所说的模糊是指这样一种现象，即某些名称运用的范围是没有严格划定界限的"②他举的例子是about a hundred（大约一百）、Sweetish（有点甜的）、greenish（带绿色的）、big（大）、Small（小）、quickly（快）、slowly（慢）等。这样，自然语言的模糊性问题又引起了人们的

---

① 转引自：《伍铁平，再论语言的模糊性》，（《语文建设》），1989年，第6期，第26页。

② 转引自：伍铁平，《论模糊理论的诞生及其研究对象与正名问题》，（《语文现代化》），1983年，第2辑，第106页。

兴趣。

英国著名哲学家和数学家罗素（B.Russell）于1923年写过一篇《论模糊性》的论文。他指出："整个语言都或多或少是模糊的"。并且举例论证了这个问题，"由于颜色构成一个连续统，因此颜色有深有浅，对于这些深浅不同的颜色，我们就拿不准是否把它们称为红色。这不是因为我们不知道'红色'这个词的意义，而是因为这个词的适用范围在本质上是不确定的。这自然也是对人变成秃子这个古老之谜的回答。假定一开始他不是秃子、他的头发一根根地脱落，最后才变成秃子。于是有人争辩说，一定有一根头发，由于这根头发的脱落，便使他变成秃子。这种说法自然是荒唐的。秃头是一个模糊概念；有一些人肯定是秃子，有一些人肯定不是秃子，而处于这两者之间的一些人，说他们必定要么是秃子，要么不是，这是不对的。排中律用于精确符号时是正确的；但是当符号是模糊的时候，排中律就不合适了。事实上，所有的符号都是模糊的。所有描述感觉特性的词，都具有'红色'这个词所具有的同样的模糊性。这种模糊性也存在于象一米或一秒钟这种表示数量的词之中，尽管这些词的模糊程度较低，而且科学曾竭尽全力使这些表示数量的词变得精确。我不会为了要使这些词变得模糊去求助于爱因斯坦。例如'一米'被定义为巴黎的一定温度下的一根测杆上两个标志之间的距离。既然这些标志不是点，而是一定大小的斑，所以它们之间的距离就不是一个精确的概念。加之温度的测量不可能超过一定程度的精确性，测杆的温度也从来不是始终如一的。基于所有以上原因，'一米'的概念是缺乏精确性的。"一秒钟"也是如此。秒是根据与地球旋转的关系下定义的。但地球不是一个刚体，且地球表面两部分转动的时间并不相同，况且所有的观测都有误差。有些事件我们可以说它们在不到一秒钟就发生了，而另一些事件则要一秒多。但是，在这两者之间必有一些事件，我们相信它们并不是都持续了同样久的时间，可是这些事件中没有一件我们能说，它们是持续了一秒多

还是少于一秒。所以，当我们说一个事件延续了一秒钟时，它的全部有价值的意义就是：不可能有精确的观察表明它持续了一秒还是少于一秒”。[①] 罗素这篇论文对传统逻辑学中的排中律提出挑战，从哲学和逻辑学上为模糊理论奠定了基础。

1933年，美国语言学家布龙菲尔德在其名著《语言论》中也指出了自然语言中存在着模糊现象。他说：“我们可以根据化学或矿物学来给矿物的名称下定义，正如我们说‘盐’这个词的一般的意义是‘氯化钠’（NaCl），我们也可以用植物学或者动物学的术语来给植物或者动物的名称下定义，可是我们没有一种准确的方法来给象‘爱’或者‘恨’这样一些词下定义，这样一些词涉及到好些还没有准确地加以分类的环境——而这些难以确定意义的词在词汇里占了绝大多数。”[②] 他进一步指出：“此外，即使我们有一些科学的（也就是普遍被承认的而又准确的）分类，我们也还往往发现语言里的意义跟这种分类并不一致。德语里把鲸鱼叫做一种‘鱼’：(Walfisch[val-'fiʃ])，而把蝙蝠叫做‘小耗子’（Fledermaus ['fle:der-'maws])。物理学家把光谱看成是不同长度的光波的连续阶程，即从4×10^{-4}m.m.到7.2×10^{-4}m.m；可是许多语言却相当任意地划分了这种阶程的不同部分而且没有确切的界限。在象紫罗兰色、兰色、绿色、黄色、橙色、红色这样一些颜色名称的意义里以及在不同语言的颜色名称里并不包含相等的差级。人们的亲属关系看来是件简单的事，可是在不同语言里所用的亲属称呼却极难分析”。[③] 后来许多学者研究颜色词的模糊性质和亲属称谓问题，正是沿着布龙菲尔德在这里所提供线索进一步深入下去的。

1937年，布莱克（M. Black）也写了一篇《论模糊》的文章，

---

① 罗素，《论模糊性》，中译文见《模糊系统与数学》，1990年，第4卷，第1期，第17—18页。

② 布龙菲尔德，《语言论》，中译本，商务印书馆，1980年，第166页。

③ 同②，第167页。

指出模糊和精确具有相对的性质，他说："绘图员画的线不论如何精确，在显微镜下看去却象一条波纹状的壕沟，离纯几何学上的理想线相去甚远"。他还指出，模糊现象指的是某种边界状况，这时不可能将某个项目划属或不划属某个范围；所有那些需要由人的感官加以辨认的东西，表示这些东西的词（如颜色词）都是模糊的①。

上述这些学者都指出了语言中的模糊现象，但是，直到1965年，美国加利福尼亚大学教授查德（L. A. Zadeh）发表了模糊集合论（fuzzy sets theory）的著名论文之后，模糊性的概念才第一次得到了完善的表示方法②。查德是一位数学家，可是，他在模糊数学方面的研究工作却首先是从观察语言符号的模糊性开始的。例如，"老年"这个概念就具有模糊性。七十岁算不算"老年"？如果算，那么，六十岁算不算"老年"？五十岁算不算"老年"？这是很难精确地回答的。查德把"老年"看成是建立在"年龄"这个论域上的一个集合，而把七十岁、六十岁、五十岁都看成是这个集合中的元素，这样，就可以研究这些元素相对于"老年"这个集合的隶属关系。这种隶属关系，很难用经典集合论中的"属于"或"不属于"某个集合的办法来描述，而可以用在多大程度上属于某个集合的办法来描述。也就是说，一个模糊集合 S的特征，是存在着一个隶属函数μ_s，对于论域U中的每一个元素x，都有一个确定的值$\mu_s(x)$，这个值刻画着元素x隶属于模糊集合的程度。③

例如，可以这样给出模糊集合"老年"的隶属函数公式：

设论域为U＝(0，150)，"老年"的隶属函数公式为

---

① 转引自：伍铁平，《论模糊理论的诞生及其研究对象的正名问题》。（《语文现代化》），1983年，第2辑，第103页。

② L. A. Zadeh, *Fuzz sets*, 《*Information and Control*》, 8(1965)，第338－353页。

③ 楼世博，孙章，陈化成，《模糊数学》，科学出版社，1983年。

$$\mu_{S老}(X)=\begin{cases}0 & x\leqslant 50,\\ \left[1+\left(\dfrac{x-50}{5}\right)^{-2}\right]^{-1} & x>50\end{cases}$$

现在把55岁代入公式计算，得到

$$\mu_{S老}(55)=\left[1+\left(\frac{55-50}{5}\right)^{-2}\right]^{-1}$$

$$=[1+1^{-2}]^{-1}=2^{-1}=0.5$$

把60岁代入公式计算，得到：

$$\mu_{S老}(60)=\left[1+\left(\frac{60-50}{5}\right)^{-2}\right]^{-1}$$

$$=[1+2^{-2}]^{-1}=\left[1+\frac{1}{4}\right]^{-1}=0.8$$

把65岁代入公式计算，得到：

$$\mu_{S老}(65)=\left[1+\left(\frac{65-50}{5}\right)^{-2}\right]^{-1}$$

$$=[1+3^{-2}]^{-1}=\left[1+\frac{1}{9}\right]^{-1}=0.9$$

采用这样的隶属函数，就可以对模糊词“老年”进行定量的描述了：五十五岁属于“老年”的程度是0.5，六十岁属于“老年”的程度是0.8，七十岁属于“老年”的程度是0.9。

在对模糊词进行定量描述的基础上，还可以把否定词“非”、连接词“或”、“与”以及程度副词“极”、“很”、“相当”、“比较”、“有点儿”、“稍微有点儿”等，也用隶属函数来加以定量的刻画。被定义了某种运算法则的否定词、连接词、程度副词，叫做模糊算子。

模糊算子的运算规则定义如下：

1.否定词“非”的隶属函数

$$\mu_{非A}=1-\mu_A$$

2.连接词“或”的隶属函数

$$\mu_{A或B}=\mu_A\vee\mu_B$$

3.连接词“与”(“且”)的隶属函数

$$\mu_{A与B}=\mu_A\wedge\mu_B$$

$$\mu_{A且B}=\mu_A\wedge\mu_B$$

4.程度副词“极”、“很”、“相当”、“比较”、“有点儿”、“稍微有点儿”的隶属函数

$$\mu_{极A}=(\mu_A)^4$$

$$\mu_{很A}=(\mu_A)^2$$

$$\mu_{相当A}=(\mu_A)^{1.25}$$

$$\mu_{比较A}=(\mu_A)^{0.75}$$

$$\mu_{有点儿A}=(\mu_A)^{0.5}$$

$$\mu_{稍微有点儿A}=(\mu_A)^{0.25}$$

例如，一位六十岁的人属于“老年”的隶属函数为0.8，那么，他属于“非老年”这一新模糊集合的隶属函数的值就为

$$\mu_{非老年}=1-\mu_{老年}=1-0.8=0.2$$

而属于“很老”的隶属函数值为

$$\mu_{很老}=(\mu_{老年})^2=(0.8)^2=0.64$$

属于“有点儿老”的隶属函数值为

$$\mu_{有点儿老}=(\mu_{老年})^{0.5}=(0.8)^{0.5}=0.9$$

再如，某人属于“高个子”的程度为0.9，而属于“胖子”的程度只有0.4，那么，他属于“高个子或者胖子”的程度为

$$0.9\vee0.4=0.9;$$

而属于“高而且胖”的程度就只有

$$0.9\wedge0.4=0.4$$

査德把普通集拓广为模糊集，为模糊数学奠定了基础，这一开创性的工作不仅拓广了普通数学的研究领域，而且开辟了在软、硬科学中提高数学适用性的广阔途径。近二十年来，模糊数学的发展非常迅速，应用相当广泛。

我们应该强调指出的是，模糊数学的产生和发展，首先是从观察研究自然语言中的各种模糊现象开始的。正如马尔可夫在对

《欧根·奥涅金》的字母序列的研究中发现了随机过程论一样，查德也是从对自然语言模糊现象的观察中发现了模糊数学的。这是语言对数学影响的生动实例。查德本人曾明确地说明："模糊集合论的这个分支的起源是从语言学方法的引入开始的，它转而又推动了模糊逻辑的发展……。在即将到来的时代，我相信近似推理和模糊逻辑将发展成为一个重要领域，从而变成研究哲学、语言学、心理学、社会学、管理科学、医学诊断、判别分析以及其它领域的新方法的基础。"① 查德这一段话，又一次说明了数学与语言之间确实存在着密切的联系。

恩格斯在《自然辩证法》一书中，早就指出了事物之间界限的不确定性，他说："一切差异都在中间阶段融合，一切对立都经过中间环节而互相过渡，对自然观的这种发展阶段来说，旧的形而上学的思维方法就不再够了。辩证法不知道什么绝对分明的和固定不变的界限，不知道什么无条件的普遍有效的'非此即彼！'，它使固定的形而上学的差异互相过渡，除了'非此即彼！'，又在适当的地方承认'亦此亦彼！'，并且使对立互为中介；辩证法是唯一的、最高度地适合于自然观的这一发展阶段的思维方法。"② 普通集A完全由其特征函数 $X_A:U\to\{0,1\}$ 刻画，它是描写"非此即彼！"的清晰概念的；而模糊集是描写模糊现象的，它容许"亦此亦彼！"的中介状态存在，因而相应的特征函数之值除了取0，1之外，还可取0与1之间的任何值，从而将特征函数推广为隶属函数$\mu:U\to[0,1]$，于是，模糊集合就可以用隶属函数μ来刻画了。显然，当隶属函数μ只取0，1两个值时，模糊集合便退化为普通集合，模糊性就变成了确定性。可见，模糊集合论是完全符合于辩证法规律的科学理论。

徐利治教授在《数学方法论选讲》一书中将数学模型分为三大

①查德，《模糊集》，中译文，载《自然科学哲学问题》，1981年，第1期，第67—68页。

②恩格斯，自然辩证法，人民出版社，1971年，第223页。

类：①

第一类是确定性数学模型，这类模型的背景对象具有确定性或固定性，对象间又具有必然的关系。

第二类是随机性数学模型，这类模型的背景对象具有或然性或随机性。

第三类是模糊性数学模型，这类模型的背景对象具有模糊性。

自然语言是一个极其复杂的符号系统，自然语言的有些规律是可以用确定性数学模型来描述的，但是，由于语言符号的随机性和模糊性，自然语言中的很多规律必须借助于随机性数学模型和模糊性数学模型，才能进行恰当的描述。

语言符号的随机性与语言符号的模糊性是两个不同的概念。

语言符号的随机性是指事件的发生与否而言，但事件本身的含义是确定的，由于条件不充分，事件的发生与否有多种可能性，在[0，1]上取值的概率分布函数就是描述这种随机性的，它经常表现为字符或单词出现概率的大小。

语言符号的模糊性是指元素对集合的隶属关系而言，事件本身的含义是不确定的，但事件发生与否可以是确定的，因而元素（事件）对集合的隶属关系是不确定的，在[0，1]上取值的隶属函数就是描写这种不确定性（即模糊性）的，它经常表现为单词含义对某一集合隶属函数值的大小。

语言符号的随机性放弃了“一因一果”的决定论，反映了“一因多果”的规律性，因此，它是由于因果律破缺而造成的一种不确定性，在用统计数学方法来描述语言时，是满足互补律的。

语言符号的模糊性摆脱了“非此即彼”的确定性，反映了“亦此亦彼”的规律性，因此，它是由于互补律破缺而造成的一种不确定性。

研究语言符号的随机性，可以把语言学的领域从必然现象扩

---

①徐利治，《数学方法论选讲》，华中工学院出版社，1983年。

大到偶然现象，研究语言符号的模糊性，可以把语言学的领域从清晰现象扩大到模糊现象。因此，语言符号随机性和模糊性的发现，都加深了我们对于语言符号本质的认识，拓广了语言学的研究领域。

除了在语言的单词含义方面存在模糊性之外，语法方面也存在着模糊性。例如汉语中单复句的划分问题，传统的划分方法是按"非此即彼"的确定性原则划分的："非单即复，非复即单，二者必居其一"。但是，由于汉语处于不断的动态变化之中，有许多句式还处于单、复句两端的中介部位，模糊性比较大，有不少句子，特别是口语句子，不能简单地归入对立的两端单句或复句中去，它们有的既象单句，又象复句，有的既不象单句，又不象复句。它们究竟象什么呢？只有把对立的*A*(单句)、*B*(复句)两端直接结合起来，重视"中介物"的"不*A*不*B*"、"亦*A*亦*B*"，不采用单句、复句二分的办法，而采用三分或多分的办法，才能使问题得到比较圆满的解决。① 此外，还有兼类词问题（如"计划、工作、编辑、出版"等算名词还是算动词)、离合词问题（如"理发、洗澡"等，也可以说"理一次发，洗一个澡"，是"理发、洗澡"分别算一个词，还是"理、发、洗、澡"分别算一个词)，看来也都应该采用模糊数学的方法来研究。

语言是约定俗成的，所以，它的明确或模糊，或者说，它的模糊实态，也应该取决于社会的规定。确定或估计模糊性的根本办法是进行社会调查。我国航天医学工程研究所通过实验调查了"快"、"慢"的语义的模糊实态。② 他们的实验是这样来进行的：

荧光屏上一次又一次的出现一个活动的光点，每次光点活动快慢不完全相同，以最慢到最快有15种均匀的级别（15个"速率元素"），编号1～15，1级最慢，2级稍快，……，15级最快。光

①陈建民，《现代汉语句型论》，语文出版社，1986年，第4页。

②龙升照等，《人机系统中人的Fuzzy概念的确定》，《模糊数学》，1981年创刊号。

点由仪器控制，随机出现，每个级别的光点都出现320次。四个青年应试者先熟悉光点运动的快慢。实验的时候，叫应试者把每个活动光点用“快”、“中”、“慢”三个概念之一判断出来。实验结果如表7.1.1所示：

表7.1.1　　“快”、“中”、“慢”调查结果

光点快慢等级	频数分布			得分分布		
	快	中	慢	快	中	慢
1	0	0	320	0	0	100
2	0	1	319	0	0	100
3	0	8	312	0	3	98
4	0	83	237	0	29	73
5	0	179	141	0	63	44
6	1	221	98	0	77	31
7	27	285	8	8	100	3
8	30	286	4	9	100	1
9	70	246	4	22	86	1
10	238	81	1	74	28	0
11	292	28	0	91	10	0
12	306	14	0	96	5	0
13	310	10	0	97	3	0
14	312	8	0	98	3	0
15	320	0	0	100	0	0

表7.1.1中，记录了实验结果的频数分布和得分分布。频数分布记录了应试者对光点快慢判断的“人次”数。例如，活动快慢为6级的光点，出现在320次中，有1人次判断为“快”，有221人次判断为“中”，有98人次判断为“慢”。得分分布是从频数分布换算而来的，以100分为满分，代表明确性最强或模糊性最弱，这一栏的数字分布相当于模糊数学中隶属函数值，只不过放大了100倍。

根据这些数据可画成如下的曲线图，如图7.1.1所示。

这个实验给我们提供了一个确定某一语义模糊实态的方法。当然，调查不一定都要用仪器来实验。也可以向有代表性的应试

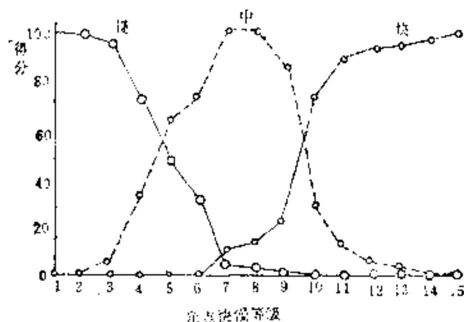


图7.1.1 “快”、“中”、“慢” 曲线图

者口头询问，统计得分，得出有关语义的模糊实态。

语言符号的模糊性是语言符号的特点，这种模糊性在很多情况下是必要的和有用的。模糊数学家哥根(J. A. Goguen)在1974年说过：“描述的不确切性并不是坏事，相反，倒是件好事，它能用较少的代价传送足够的信息，并能对复杂事物作出高效率的判断和处理。也就是说，不确切有助于提高效率”。①他还说：“我们必须至少在语言交际方面放弃这样一种观念：‘较准确’总是较好”。其实不然，“模糊不仅对人类来说比较适合，……对机器来说，实际上也更为有效”。②德国数学家弗雷格（G. Frege）把模糊性看作是人的一种直觉（intuitive feeling）。法国学者杜梅（M. Dummett）说：“模糊性是人类语言不可或缺的一个特点。如果人类语言的许多表达不显现模糊性，我们就不能象现在我们这样运用我们的语言……任何对模糊性的令人满意的解释必须至少能

①转引自Negoita等著，《模糊集在系统分析中的应用》，中译本，1980年版，引言。

②J. A. Goguen，《论模糊机器人的设计》，见查德等主编《模糊集及其在认识过程和决策过程中的应用》，美国学术出版社，1975年，第430—445页。

说明人的这种直觉"。① 意大利学者泰尔米尼（S. Termini）说："不精确性在科学中所起的重要作用已为人所周知，例如。路德维希1981年就说过，大学生从开始学习起就必须认识到的有关实验物理学的最基本的事实之一是：没有一种测量是精确的……一个理论物理学家应该认识到：作为客观现实图景的任何一种数学理论都不能看作是一副精确的图画。他应该认识到，在所谓精确理论同客观现实的近似图画之间原则上并没有区别。"②

语言中有许多说法就是为了减少明确性、增加模糊性的。例如"大约、差不多、基本上、大概、上下、左右、光景、也许、多半、在一定程度上……"等，这些词语人人都要用，这说明了模糊常常是必要的、有用的。在许多情况下，避免明确可以使语言更加含蓄婉转，留有余地。③

"弄、搞"等动词在汉语中很常用，它们的含义很笼统，什么动作都能表示，用起来不假思索，方便灵活。叫人"弄点吃的来"，他可以酌情选用最方便的途径去弄，如果明确地指定他是做、是买、是要、是拿、是借、是炒、是蒸、是煮，他就受到了限制。模糊性给他留下了很大的选择余地。

有意识的模糊化常常是语言中采用的手段之一。比喻、影射、暗示、隐讳、欲说又止等增加模糊性的手段常常会取得很好的效果。

所以，有人认为描述的精密和实用是有矛盾的。贝尔曼(R. E. Bellman) 在1973年说过："要想确切地描述任何现实的物理状态，事实上是办不到的。这是一个公认的并经过检验的事实。因此，描述（对于通讯、作决定，推而广之对于人的一切活动都是不可少的)的主要问题便是：减少必然会有的不确切性，使它达到无关紧要的程度。为了把整个问题描述得详尽，我们必须在准确

①转引自伍铁平，《从语言的模糊性谈到人脑与电脑的区别》，1989年。
②同①。
③刘泽先，《语义的模糊和明确》，(《语文现代化》)，第9辑，1990年。

和简明之间取得平衡，既减少复杂性而又不过于简单化。"[①] 这就是所谓的"互克性原理"。

当然，在科学技术领域中，科技术语要求明确地表达出有关的技术内容。随着科学技术的发展，人们对于某些被认为是模糊的术语会描述得越来越精确。

例如，颜色词一般被认为是典型的有模糊语义的词。因为在"红—橙—黄—绿—蓝—紫"的颜色系列中，"红"与"橙"之间，"橙"与"黄"之间，……等等，都没有明确的边界。《现代汉语词典》把"绿"解释为"象草和树叶茂盛时的颜色"，《辞海》则把"绿"规定为"青中带黄的颜色"，由于草和树叶的颜色既不完全相同而且时有变化，青中带黄的程度也不甚明确，"绿"的语义显然是十分模糊的。

但是，由于科学技术的发展，当人们对颜色这一事物的认识深入到数量界限的程度，能够准确地从数的角度来描述各种颜色的时候，这些颜色词的模糊性也就逐渐消失了。

现代科学把"颜色"定义为视觉的基本特征，是不同波长的可见光引起的视觉器官的不同感觉，并且根据可见光的不同波长明确地划分了红、橙、黄、绿、蓝、紫的界限：

红：波长为 0.77—0.622微米的可见光引起的人的颜色感觉。

橙：波长为 0.622－0.597 微米的可见光引起的人的颜色感觉。

黄：波长为 0.597—0.577 微米的可见光引起的人的颜色感觉。

绿：波长为 0.577—0.492 微米的可见光引起的人的颜色感觉。

蓝：波长为0.492—0.455微米的可见光引起的人的颜色感觉。

---

① 转引自C. V. Negoita等著，《模糊集在系统分析中的应用》，中译本，1980年版，引言。

紫：波长为 0.455—0.390 微米的可见光引起的人的颜色感觉。

数量是任何事物固有的规定性，人对事物的认识只有深入到事物的数量才是真正的深化。因此，一切语义模糊实质都是数的模糊，而具体的模糊词是一个历史范畴，随着人们对事物的数的认识的发展，原来的模糊词有可能成为精确词。

现代科学标志着人类整体的认识世界对客观世界所达到的最新认识程度，但个体的认识世界却并不都与之一致。不同文化程度、不同专业的个体都不能同样地达到这样的认识程度。尽管科学家对于“绿”作了上述严格的规定，但是，当人们在遇到似绿非绿、绿中带黄或绿中带蓝的颜色时，并不都能根据可见光的波长来决定它究竟是否为“绿”，因此，就人类的个体而言，颜色词在个体认识上的模糊性仍然是存在的。①

又如，对于“胖”、“瘦”这样的词，其语义也是模糊的，日常语言中究竟什么算“胖”，什么算“瘦”，并没有一个明确的标准，可是，在医学上却有一个简单的公式作为判断胖瘦的标准：

身高厘米数－105＝标准体重公斤数

如果你的体重超过标准体重，那就算“胖”，如果低于标准体重，那就算瘦。这样，“胖”、“瘦”这样的模糊词的界限也就明确起来了。

再如，针麻手术时病人的感觉，究竟到什么程度算“疼”，其界限也是不明确的。但是，医学上要求用明确的语言来说明疼痛的程度，作了如下的规定②：

“0”无痛。

“＋”轻痛。病人表现皱眉、呻吟、微汗、血压波动在20毫米

---

①符达维，《模糊语义问题辨述》，《中国语文》，1990年，第2期，第109页。

②粟载福、欧阳绵，《可能性理论在针麻手术效果评级中的应用》，《模糊数学》，1981年第2期。

汞柱以内，脉搏波动在20次/分以内。

"++"中痛。病人表现屏气、呼痛、握拳、四肢挪动、出汗较多，血压波动在20～30毫米汞柱之间，脉搏波动在20～30次/分之间，或血压骤降20毫米汞柱，脉搏骤降20次/分之内。

"+++"剧痛。病人表现为咬牙、大声连续呼痛、大汗、躁动不合作，血压波动在30毫米汞柱以上，脉搏波动在30次/分以上。

这样，"无痛"、"轻痛"、"中痛"、"剧痛"等模糊词的语义也就有了比较明确的界限。

罗素在《论模糊性》中曾指出了"一米"、"一秒"也是模糊的。近年来，人们对于"米"、"秒"等基本计量单位有了新的规定。1960年10月，第十一届国际计量大会上通过了新的决议，规定一米等于氪86在真空中在 $2P_{10}$ 和 $5d_5$ 两个能级之间跃迁时所发射的橙色光波波长的1 650 763.73倍。这样，罗素对于"一米"的模糊性的论述就失去了根据。同样，国际计量大会对于"一秒"也作了新的规定，一秒等于铯138原子基态的两个超精细能级之间跃迁时所吸收或放出的电磁波周期的9 192 631 770倍。这样，罗素对于"一秒"的模糊性的论述也就站不住脚了。

可见，模糊词的含义会随着历史的变化和科学技术的发展而发生变化，它们并不是永恒不变的。随着科学的发展，一些模糊词的模糊语义会逐渐消失。但是，由于世界是无限的，真理是不可穷尽的，所以，人们对于客观世界的认识是没有止境的，人们对于事物之间界限的认识也是没有止境的，一些模糊词的模糊语义消失了，还会产生一些新的模糊词，而且，就是被认为已经消失了的模糊语义在新的认识水平和科学层次上看来，又可能会成为在新的意义上的模糊语义，模糊词将永远存在，而且永远在变化着。模糊词是一个历史范畴，模糊语义并不是永恒不变的。从这个意义上说，模糊语义的研究不仅对于语言学本身还是对于整个现代科学的发展，都是很有价值的。在研究语言符号的模糊

性的基础上而产生的模糊数学，正在不断完善它的基本理论，不断拓广它的应用领域。现在，模糊数学的应用已涉及到聚类分析、图象识别、自动控制、机械故障诊断、系统评价、数据结构、情报检索、机器人、人工智能、逻辑等许多领域，同时，模糊数学又反过来应用于语言学。模糊语言的研究已引起了语言学家们的浓厚兴趣。1972年，在美国纽约举行的一次词典学国际讨论会上，美国语言学家雷柯夫（G. Lakoff）作了一个在词汇研究方面应用模糊数学的报告。雷柯夫高兴地说："我们现在有了一个'可爱的术语'——模糊集合"。他在讨论会结束时又指出，模糊性将成为语言学的一个主要的研究领域①。

第2节　模糊数学在语言研究中的应用

近年来，模糊数学在汉语研究中得到了广泛的应用。许多学者在词典学、词源学、修辞学、术语学、方言学中，都使用模糊数学的方法，取得了一定的成果。这里不可能面面俱到地介绍这些研究成果，只是通过一些实例来说明模糊数学与语言学的关系。

1848年，德国语言学家格里木（J. Grimm）出版了一本书，叫做《德意志语言史》，认为德国方言不是高地德意志语，就是低地德意志语。在他看来，属于既不是高地德意志语、又不是低地德意志语的法兰克方言（一种德语方言）早已完全消失。恩格斯深入研究了法兰克方言，认为这种方言是一种既是高地德意志的又是低地德意志的方言，也就是说，法兰克方言是一种"亦此亦彼"的中介物②。

---

①楼世博、孙章、陈化成，《模糊数学》，科学出版社，1983年，第75页。
②恩格斯，《自然辩证法》，人民出版社，1971年版，第195页，书中把格里木译为格林。

在恩格斯研究法兰克方言的启示下，我们采用模糊数学来构造方言的数学模型。①

我们发现，语言符号的离散性和模糊性具体体现在方言的分布上，就是方言既不连续（离散性）而又相互交错（模糊性）这两个特点。

方言不连续的例子很多，例如，北京话与哈尔滨话很接近，而介于这两者之间的辽宁话，却与北京话相去甚远。又如，离成都较远的重庆，古入声字都变为阳平，在成都西面的荣庆则保存了入声，在成都、重庆之间的隆昌、内江等地，古入声字则又变为去声。再如，杭州附近都是属于吴语区的方言，而杭州一带却属于官话区，形成了一个语言岛。这在外国方言中也不乏其例。例如，在方言繁多的日本，日语词“借りぬ”(かりぬ),在本州关东、东京地区，都念作Kariru，因而成了其标准音。但是，在离东京很近的横滨附近，这个词却念Kain，相去甚远。而且，在本州中部一大片念Kariru的地区中，却又散布着一些离散的点，这个词念作Kareru。有趣的是，这个词的读音沿着南起太平洋伊势湾、北到日本海富士湾的一条曲线而一分为二，曲线以北念作Kariru，而曲线以西则读作Karu。所有这些现象,都说明方言读音的变化是不连续的。至于以单个形态出现的词汇和句法特点，当然就更不是连续的了。

再谈方言分布相互交错的特点。例如，以广大的吴语区内部而言，江苏靖江附近方言的分布模式十分有趣。靖江县说吴语，靖江县北面说苏北话，南面说官话，形成一个语言岛。而且在靖江县内部，又有一个小区域讲官话，形成“岛中之岛”。在江苏的启东、海门附近，有一种当地老百姓称为“夹沙江”的有趣现象，指的是沙里话（属吴语区的海门话）和江北话（官话）往往

①钱锋、冯志伟，《试论模糊数学在方言研究中的应用》，《华东师范大学学报》(哲学社会科学版)，1983年，第4期。

是以一个村一个村夹杂着分布的。这些现象说明，方言分布的界限不是很清楚的，存在着模糊性。

因此，我们可以采用模糊数学的方法来描述方言。

对于模糊集合，不能象经典集合那样画出有边界的示意图。我们可以用一种类似于地理“等高线”的图来表示模糊集合，叫做α截集合（α level set）。

假定A是某一论域上的模糊集合，那么，A的α截集合可定义为：

$$A_\alpha=\{u|\mu_A(u)\geqslant\alpha\},\quad 0\leqslant\alpha\leqslant 1$$

也就是相对A的隶属函数值不小于α这个确定值的元素x的集合。A_α当然是一个非模糊集合，因为对任意的x，都可以根据

$$\mu_A(x)<\alpha$$

或是

$$\mu_A(x)\geqslant\alpha$$

来判定

$$x\notin A_\alpha$$

或是

$$x\in A_\alpha$$

于是，一个模糊集合就可以看成是所有这种A_α所组成的了，也就是说，A有了新的表达式：

$$A=\int_0^1\alpha A_\alpha$$

当α是离散值时，也可以写成

$$A=\sum_\alpha \alpha A_\alpha$$

有时候，我们只对某一些离散的α值感兴趣，就可以选取一组

$$\{\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_k\}\quad 其中\alpha_1\geqslant\alpha_2\geqslant\cdots\geqslant\alpha_k$$

而研究

$$A_{\alpha_i}=\{u|\mu_A(u)\geqslant\alpha_i\}\quad i=1,\ 2,\ \cdots,\ k$$

或者说一组

$$\{A_{\alpha_1}, A_{\alpha_2}, \cdots, A_{\alpha_K}\}$$

于是，模糊集合A就可以用图7.2.1来表示了

下面我们会看出，这种图对于绘制方言地图很有用处。

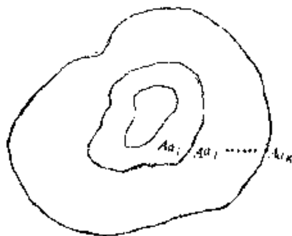


图7.2.1　模糊集合的一种图示

如果我们要研究同一论域上的两个模糊集合之间的差异，可引入模糊集合A和模糊集合B之间的距离（distance）的概念。

假设刻画A、B的隶属函数分别为μ_A和μ_B，那么，A与B之间的距离定义为

$$d_p(A,B)=\left(\sum_u |\mu_A(u)-\mu_B(u)|^P\right)^{1/p}$$

在一般情况下，取$P=2$，于是有

$$d_2(A,B)=\sqrt{\sum(\mu_A(u)-\mu_B(u))^2}$$

这叫做A与B的欧几里得距离。

现在，我们根据上面的论述来研究如何统计方言的差别，如何在统计的基础上绘制方言地图。

1.点方言特点的数学描述

要考察某一方言点x上的方言（称之为点方言x）相对于某一方言S的隶属关系，必须先就两者的语音、词汇和语法做大量的调查工作，我们可以把点方言看成是由语音、词汇和语法三个分量所组成的向量，即

$$S=(S_{语音}, S_{词汇}, S_{语法})$$

或者

$$S=(S_1, S_2, S_3)$$

相应地，方言点x也有

$$x=(x_1, x_2, x_3)$$

各个分量相应于S的各个分量都可以分别由统计结果计算其隶属

函数

$\mu_{si}(x_i)$　　　　　　$i=1, 2, 3$

假设我们来求$\mu_{s_1}(x_1)$。其中，x_1 即点方言x的语音特点，它也是一个向量。例如

$x_1=(x_{11}, x_{12}, x_{13})$

其分量分别表示声母特点、韵母特点和声调特点。

同样，方言S也是如此：

$S_1=(S_{11}, S_{12}, S_{13})$

所有这些S_{ij}，x_{ij}都是模糊集合。

点方言x语音部分相应方言S的隶属函数可以这样求得：

$$\mu_{s_1}(x_1)=\sum_{i=1}^{3}K_i\cdot\mu_{s_{1i}}(x_{1i})$$

其中，$\mu_{s_{1i}}(i=1, 2, 3)$表示的是：点方言的x_{11}，x_{12}和 x_{13}看作元素分别相对模糊集合$S_{1i}(i=1, 2, 3)$ 的隶属函数 的值。这一函数值当然可以通过对于两者各自的调查、统计、分析和比较而估计出来。式中$K_i(i=1, 2, 3)$称为“权”，它是满足

$$\sum_{i=1}^{3}K_i=1, \qquad K_i\geqslant 0 \ (i=1, 2, 3)$$

的一组常量，它们分别表示声、韵、调三者在形成方言的整体语音特色中所起的作用，也即比重。这个权是语言学家研究的结果。例如，可以取

$K_1=0.35$，$K_2=0.4$，$K_3=0.25$

也可以根据研究结果取其它的数值。

同理，可以求得

$$\mu_{S_2}(x_2)=\sum_{i=1}^{3}L_i\cdot\mu_{s_{2i}}(x_{2i})$$

$$\mu_{s_3}(x_3)=\sum_{i=1}^{3}M_i\cdot\mu_{s_{3i}}(x_{3i})$$

并最终推出

$$\mu_S(x)=\sum_{i=1}^{8}\rho_i\mu_{si}(x_i)$$

其中

$$\sum_{i=1}^{8}\rho_i=1$$

也是权，它表明语音、词汇和语法三者在形成方言对比中所占的比重。

根据上面关于 $\mu_s(x)$ 推导过程的简单讨论，我们有理由把 $\mu_s(x)$ 作为描述点方言 x 与方言 S 之关系的一个全面的和数量的标志。

2.方言差异的数学描述

我们可以用欧几里得距离的概念来描述方言之间的差异程度。首先，我们建立几个大方言 S_i（如吴方言区、湘方言区、北方方言区等）相对普通话 P 的隶属函数

$$\mu_P(S_i) \qquad i=1,\ 2,\ 3,\ \cdots,\ n$$

这样，我们就有可能求出各大方言之间的距离：

$$d(S_i,S_j)=\sqrt{\sum_{k=1}^{8}[\mu_{P_k}(S_{iK})-\mu_{P_k}(S_{jK})]^2}$$

$$i,\ j=1,\ 2,\ 3,\ \cdots,\ n$$

其中 $\mu_{P_k}(S_{iK})$ 和 $\mu_{P_k}(S_{jK})$（$K=1,\ 2,\ 3$）分别表示大方言 S_i 和 S_j 的语音、词汇和语法相对普通话 P 的语音、词汇和语法的隶属函数。

仿照上式，我们也可以建立一个大方言中两个点方言 x，y 之间的差异，即

$$d(x,y)=\sqrt{\sum_{i=1}^{8}[\mu_{si}(x_i)-\mu_{si}(y_i)]^2}$$

3.方言地图

一般的方言地图大都是描绘单个语音、词汇现象在地理上的分布情况，不能给人们带来一种总体的和数量的概念。

我们建议，采用α截集合的概念来绘制方言地图。

以吴语 W 为例，我们取吴语的中心地带为中心，取一组数值：

$$\{\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_k\}$$

例如，$\alpha_1=0.95$，$\alpha_2=0.90$，$\alpha_3=0.85$，…，$\alpha_{19}=0.05$等等。

然后把符合

$$W_{\alpha_i}=\{u|\mu_W(u)\geqslant\alpha_i\}$$

的各方言点μ连接起来，就能形成如7.2.2的图。

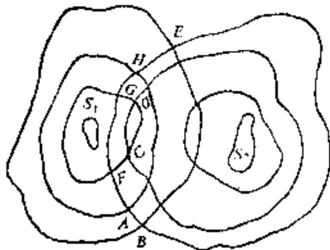


图7.2.2　模糊集合指导下的方言地图

从这张地图上，我们可以直观地看出几种相邻方言之间在数量上的相互关系以及它们与某一大方言关系的深浅程度。借助于这种方言地图的数量性和整体性，我们还可以观察到一些有趣的现象。例如，在图 7.2.2中，有两个大方言S_1和S_2，图中显示了这两个方言边缘地带的情景。在这个边缘地带有 8 个点方言，即A，B，C，D，E，F，G和H，其中

$$\mu_{S_1\cap S_2}(F)=\mu_{S_1\cap S_2}(G)=0.5$$

而其它的点方言

$$\mu_{S_1\cap S_2}(A)=\mu_{S_1\cap S_2}(B)=\cdots=\mu_{S_1\cap S_2}(H)=0.25$$

这就说明了F和G两个方言点相对于S_1和S_2的共同特点而言最为接近，而其它各方言，不管它们与S_1和S_2的距离如何，都与S_1和S_2的共同点相差较远。所以，从图上我们很快就能找到那些在最大程度上接近某两种或三种大方言的点方言。

目前，机器系统和人文系统的控制都存在着智能化的倾向，计算机一体化的加工系统则要求一系列的人工智能技术。人们普

遍认为，智能化的决策管理系统是第二次电子革命的一个极为重要的方面，也是未来10年科学技术的主要挑战之一。

在人工智能研究中，自然语言的表达和理解技术是一个十分困难的问题。科学家们已经认识到，这个问题比他们原来所预料的更加艰难，美国国会技术评价办公室最近指出，要使计算机具备一个5岁小孩的自然语言理解能力说不定是20年之后的事。

自然语言的表达与理解的主要困难在于自然语言本身的模糊性。这种困难的内在原因是我们对于人类如何贮存和处理模糊信息的机制还不十分清楚，外在原因是我们还没有一种适合于处理模糊信息的工具。

由模糊数学创始人查德亲自开拓的可能性理论、模糊语言方法以及由此而产生的模糊语言逻辑、自然语言意义表达和近似推理已构成一个知识分支，正在把克服上述自然语言理解和表达技术中的困难当作自身的研究目标，目前已取得了一些令人鼓舞的成果。①

我们在前面讲过的乔姆斯基的形式语言理论、库拉金娜的语言集合论模型以及各种逻辑语法都很难反映自然语言中的模糊性，而实际上自然语言的各个方面所表现出来的特性几乎都只能用程度来描述，语义本身有程度问题，是否符合语法也有程度问题，因此，一个没有过渡的真值或隶属程度的逻辑系统（如二值逻辑系统），不能用来作为自然语言理解的逻辑系统，这种逻辑系统对于意义表达、知识表达和信念强度的表达是无能为力的。

查德在1981年提出的"测分语义学"（Test-Score Semantics）为解决这样的问题提供了有用的工具②。测分语义学认为，一个语言实体（一个谓词、一个命题、一个问句或一个命令）的作用一

---

①陈国权，《知识工程中自然语义的模糊表达》，科学出版社，1989年。

②L. A. Zadeh, *Test-Score Semantics for natural language and meaning representation viz PRUF*, Technical Note 247, Univ. of California, 1981.

般而言在于引导论域中目标或关系的集合上的弹性限制。因此，这样一个语言实体的意义可以定义为一个过程；这个过程由三个部分组成：

①识别由语言实体引导的限制；

②描述为确定每种限制满足程度必须进行的试验；

③规定部分试验所得的测分如何合成以产生总测分。

从这个角度来看，自然语言中语言实体的意义，就是对所论语言实体中包含的弹性限制的试验。

如果我们想试验一个有语言能力的机器人是不是理解一个命题的意义，那么，就可以用测分语义学的方法来试验。

设试验对象为H，所论的命题为p

$p \triangleq$张三正在和李四一起跳舞

其中，符号"$\triangleq$"表示“按定义等于”之意。

最简单的方法是让H观看各种各样关于张三和李四联合活动的情景W(例如，一些照片)，请H给每个情景 $w \in W$ 与他所感知的p 的意义的符合程度打分，得分记为 $c(w)$。如果H对每个 w都能够给出正确的结果，即H能通过这个试验，那么，就可断定H明白p的意义。如果H还能够说清楚他在W上做试验获得$c(w)$ 的试验过程，那么，就说明H不但在直觉上了解了p的意义，而且把对p的意义的理解提高到了理性的高度。由此可以看出，正是H所说明的试验过程而不是别的什么东西代表了命题p的意义。

在测分语义学中，$c(w)$是直线上或半序集上的一个点，通常取单位间隔[0,1]作为$c(w)$的值域。$c(w)$还可以是单位间隔上的概率分布或可能性分布，甚至是概率分布与可能性分布的组合。

除了用情景作实验，语义试验还可以在另一个更抽象的层次上进行。假设我们事先有一组描述情景或实际状态的特征，或者事先已建立了一个关系数据库$\mathscr{D}$。在这种情况下，可以把命题p出示给H，让H在$\mathscr{D}$上做一组试验T以产生测分τ，写成公式，有

$$\tau = T(\mathscr{D}) = \mathrm{Comp}(p,\ \mathscr{D})$$

式中，试验T可以看作命题p的意义表达；测分τ是p与$\mathscr{D}$的符合程度的测度。

一般地说，试验T是由若干个分试验T_1，…，T_n组成的，总测分τ是分试验测分τ_1，…，τ_n的组合。其中，τ_i，$i=1$，…，n分别是T_i的测分。在测分语义学中，根据情况的不同，总测分可以是间隔[0，1]之中的一个数值，也可以是一个矢量，$\tau=(\tau_1,\cdots,\tau_r)$，这个矢量中的每个分量是间隔[0,1]中的一个数、一个概率分布或一个可能性分布。

测分语义试验是在数据库上进行的。数据库由许多关系组成，每个关系由表格来表示。一个表格有其栏头和表列量。栏头包括关系的名称和变量的名称，表列量则列出变量的值，即数据。例如，表7.2.1就是一个这样的数据库

表7.2.1　　　　数　据　库

病　　人	姓　　名	年　　龄	身　　高
	张　　三	25	170
	李　　四	30	165
	王　　五	46	175
	⋮	⋮	⋮

在这个数据库中，"病人"是关系的名称，"姓名、年龄、身高"是关系的三个变量，表列量如"李四、30、165"分别是变量"姓名、年龄、身高"的一个值，即数据。在分析中，当暂时不用具体数据时，这个关系可用其栏头来代表，写为

病人｜姓名｜年龄｜身高｜

或写为

病人〔姓名；年龄；身高〕

显然，这是一个普通关系。

当讨论模糊关系时，必须引入隶属度。模糊关系数据库的一般形式如表7.2.2所示。

表7.2.2　　　　　　　　模糊关系数据库

R	X_1	X_2	…	X_n	μ
	⋮	⋮	…	⋮	⋮
	r_{i1}	r_{i2}	…	r_{in}	μ_i
	⋮	⋮	…	⋮	⋮

在这个模糊关系数据库中，r_{i_k}，$k=1$，…，n是第i行、X_k列的表列数据，而μ_i是n重组$r_i \triangleq (r_{i_1}, \cdots, r_{i_n})$在模糊关系$R$中的隶属度，为了方便，有时$r_i$写为$(r_{i_1}, \cdots, r_{i_n})$。

例如，设模糊关系为"胖"，其变量为身高和体重，身高以厘米为单位，体重以公斤为单位，那么，我们可以有

表7.2.3　　　　　　　　模糊关系"胖"的数据库

胖	身　高	体　重	μ
	155	70	1
	168	63	0.8
	180	70	0.5

这个关系数据库的第二行中表列数据表示身高为168厘米、体重为63公斤的人定义为"胖"的程度是0.8。

在自然语言中，模糊词"许多、大多数、几个、少许，几乎"等引导的有弹性限制的试验要以基数来表示。

模糊集合的基数的概念是一般集合元素数概念的推广。

设A是模糊集合，

$$A=\mu_1/u_1+\cdots+\mu_n/u_n$$

式中，u_i，$(i=1, \cdots, n)$是论域U中的元素。

我们把隶属度的算术和取为模糊集合A的基数，求得和可进行舍入，只取其最接近的整数。这种基数称为A的Σ计数，或非模糊基数，记为$\Sigma \mathrm{count}(A)$，

$$\Sigma \mathrm{Count}(A) \triangleq \sum_{i=1}^{n} \mu_i$$

例如，$\Sigma \mathrm{Count}(1/a+0.8/b+0.5/c)=2$。

下面，我们举例说明表达命题意义的试验的全过程。①

表达命题的意义可以体现为如下的过程：

①在关系数据库中选择作试验的关系；

②决定作什么性质的试验；

③决定各个分试验所获得的测分如何结合以求得总测分。

在关系的选择方面，必须考虑到接受者的知识状态，因为命题的意义表达受到人们对同这个命题有关的概念和变量的感知的影响。我们假定，对于接受者说来（不论这个接受者是智能机器人还是普通人），这些概念和变量的意义是已知的。"已知"意味着接受者认识命题中出现的基本记号以及这些记号所代表的概念及属性。当然，在测分语义学中，记号、概念与属性之间的对应可以是有弹性的。

我们来考虑如下的命题，

$p \triangleq$过量饮食导致肥胖症

假定这一命题被理解为

$q \triangleq$大多数饮食过量的人都肥胖

并且假定，意义表达过程的接受者知道"大多数、过量饮食、肥胖"等词汇的含义。在这种情况下，数据库应该包含的关系为

$\mathscr{D}1 \triangleq$人物[姓名；年龄；体重；身高；食量]

+肥胖[年龄；身高；体重；μ]

+饮食过量[食量；μ]

+大多数[r；μ]

式中，"$\mathscr{D}$"代表数据库中所有与给出的命题有关的关系，"+"代

---

①本例引自：陈国权，《知识工程中自然语义的模糊表达》，科学出版社，1989年，第66页。

表并运算。

数据库中的第一个关系为“人物”,此关系记录了被研究的人的姓名、年龄、体重、身高、食量的统计数据。其中，食量这一项可以用实际消耗食物数与该年龄、体重、身高的人的正常消耗量之比来表示。

数据库中的第二个关系为“肥胖”,此关系定义一个有一定年龄、身高、体重的人属于肥胖之人的模糊集合的程度。这个隶属度通常用符号μ表示，μ是年龄、身高和体重的函数。

数据库中的第三个关系为“过量饮食”,此关系定义了有一定食量的人属于过量饮食之人的模糊集合的程度。隶属度也用μ表示。

数据库中的最后一个关系“大多数”把模糊词大多数定义为单位间隔中的模糊集合，r代表一个数值比例,如$r=0.7$，$\mu=0.8$表示70％这个比例算为“大多数”的隶属度为0.8。

上述数据库$\mathscr{D}1$还可以简化为如下的数据库$\mathscr{D}2$。

$\mathscr{D}2$≙人物［姓名］+肥胖［姓名；μ］+过量饭食［姓名；μ］+大多数［r；μ］

在数据库$\mathscr{D}2$中,关系“肥胖”和“过量饮食”都直接定义在人物之上,而不通过年龄、体重和身高这些中介变量的数值来定义。可以预料，从数据库$\mathscr{D}1$作实验所得到的意义表达比从数据库$\mathscr{D}2$作实验所得到的意义表达更为透彻。

我们在数据库$\mathscr{D}2$上描述命题p的意义表达的全过程，这将是一个试验命题p与数据库$\mathscr{D}2$之间兼容性检验的过程。分如下4步：

第1步：计算“人物”中过量饮食的人数。用“姓名i”记“人物”中第i个人的姓名，根据Σcount表达式，有

$$\sum \text{Count}(\text{过量饮食}) = \sum_i ({}_\mu\text{过量饮食}[\text{姓名}=\text{姓名}i])$$

第2步：计算“人物”中过量饮食的胖子的人数。这时，要计算模

糊集合“过量饮食”和“肥胖”的交集的基数$\sum$Count。姓名i在这一交集中的隶属度由下式给出：

$$\mu_{\text{过量饮食}\cap\text{肥胖}}(\text{姓名}i)=\mu_{\text{过量饮食}}(\text{姓名}i)\wedge\mu_{\text{肥胖}}(\text{姓名}i)$$

其中

$$\mu_{\text{过量饮食}}(\text{姓名}i)={}_{\mu}\text{过量饮食}[\text{姓名}=\text{姓名}i]$$

$$\mu_{\text{肥胖}}(\text{姓名}i)={}_{\mu}\text{肥胖}[\text{姓名}=\text{姓名}i]$$

结果，过量饮食的胖子的$\sum$count为

$$\sum\text{Count}(\text{过量饮食}\cap\text{肥胖})$$

$$=\sum_i\Big(({}_{\mu}\text{过量饮食}[\text{姓名}=\text{姓名}i])\wedge({}_{\mu}\text{肥胖}[\text{姓名}=\text{姓名}i])\Big)$$

第3步：计算过量饮食的胖子在过量饮食的人中的比例。

$$r\triangleq\frac{\sum\text{Count}(\text{过量饮食}\cap\text{肥胖})}{\sum\text{Count}(\text{过量饮食})}$$

$$=\frac{\sum_i\Big(({}_{\mu}\text{过量饮食}[\text{姓名}=\text{姓名}i])\wedge({}_{\mu}\text{肥胖}[\text{姓名}=\text{姓名}i])\Big)}{\sum_i\Big({}_{\mu}\text{过量饮食}[\text{姓名}=\text{姓名}i]\Big)}$$

第四步：上式表示的比例值r满足由模糊词“大多数”引导的限制的程度。这个满足程度为

$$\tau={}_{\mu}\text{大多数}[r=\gamma]$$

这个满足程度是一个测分，可以解释为在给定$\mathscr{D}$之下p的真实值或在给定p之下$\mathscr{D}$的可能值。

从这个例子中，我们可以得出几点对测分语义学有普遍意义的结论：

①命题p的意义由产生测分τ的试验的全过程所表达。

②试验的描述只涉及数据库中关系表格的栏头，与数据本身

无关，因此，试验所表达的是p的内涵。

③试验的结构取决于关系栏头的选择，因此，若用另一个数据库（比如$\mathscr{D}1$)进行试验，试验过程的描述将不同。而且，即使关系表格的栏头相同，若采用不同的基数定义，试验过程的描述也有所不同。

④在测分语义学中 $\mathscr{D}$ 的选择会影响意义表达的深度。一般说来，$\mathscr{D}$的详细程度决定了意义表达的深度。例如，$\mathscr{D}2$比$\mathscr{D}1$ 的详细程度低，相应的试验过程传达的关于p的意义的信息就比较少。

对于命题

$p\triangleq$过量饮食导致肥胖症

最简单的数据库$\mathscr{D}$为

$\mathscr{D}\triangleq$因果[原因；结果]

其中，"因果"是一种关系，因果关系中列出了各种原因及其引起的结果。对于这样一个数据库，试验简化为决定下面条件是否满足：

(过量饮食，肥胖)$\subset$因果

这个条件说明，二元组（过重饮食，肥胖）是因果关系中的一个元素。在形式上，这个式子和一般语义网络中p的意义表达相同。

数据库$\mathscr{D}$的分辨能力有限。试比较如下两个命题的表达。

设命题

$p\triangleq$过量饮食导致肥胖症

$p'\triangleq$肥胖症归因于过量饮食

根据因果关系数据库$\mathscr{D}$的定义，可知p和p'的意义相同。

但是，假如我们把p'理解为q'

$q'\triangleq$大多数肥胖的人饮食过量

并在前面定义的数据库$\mathscr{D}2$上做试验，那么，可求得q'的测分为：

$$\tau = {}_{\text{r}}\text{大多数}\left[r = \frac{\sum \text{Count}(\text{肥胖} \cap \text{过量饮食})}{\sum \text{Count}(\text{肥胖})} \right]$$

这个结果与在数据库𝒟2上求得的对p的测分不同，q'的τ中，s的分母为$\sum$Count(肥胖)，p的τ中，r的分母为$\sum$count（饮食过量），这个差别正好反映了p'和p的差别。

除了测分语义学之外，查德还提出了通用可能性模糊关系语言（Possibilistic Relational Universal Fuzzy Language，简称PRUF），利用PRUF的翻译规则，可把自然语言中的表达式翻译为PRUF的表达式，从而实现了自然语言语义的模糊表达。限于篇幅，这里就不再详述了。但是，由以上叙述我们已可以看出，由于语言符号具有模糊性，模糊数学的发展和完善为数理语言学深入研究语言的模糊现象提供了有力的武器。

在本书中，我们从语言符号随机性、冗余性、离散性、递归性、层次性、非单元性、模糊性等七个方面，论述了数学与语言之间的相互关系，说明了数学已经深入到了语言研究的各个领域。查德在《模糊语言、语言变量及模糊逻辑》一书中说得好："一种现象，在能用定量的方法表征它之前，不能认为已被彻底地理解，这是现代科学的基本信条之一。开耳芬（W. Thomson，1892年封为Lord Kelvin）在1883年说过：'在物理科学中，研究任何论题的关键的第一步是寻找它的数值计算原理和与之有关的一些性质的测量方法。我常说，要懂得一点东西，你就必须设法把这件东西测量出来并且把它表达为数字。相反，当你不能把它测量出来又无法把它表达为数字时，你对这件东西的知识是贫乏而不充分的，知识可能在你的头脑中，但无论如何，你的思想还未进入科学的境界'。"① 德国著名哲学家康德（I. Kant）曾经表述过这样一种思想："任何科学含有数学成分的多寡决定了它在多大程度上够得上成为一门科学"。② 美国语言学家华特茂（J. Whatmough）

---

① 查德，《模糊集合、语言变量及模糊逻辑》，中译本，科学出版社，1982年。

② 转引自舒哈特（H. Schuchadt）的论文《*Sachen und Worter*》（物与词），载于《舒哈特语言学论文选》（俄译本，莫斯科，1970年，第275页）。

在第八届国际语言学家大会的发言中指出："有一种先入之见，认为语言学、物理学、生理学和神经学都彼此毫无关系，正是这种认识阻碍了而且仍然在阻碍着进步，在语言学中尤其是这样，但是，只有数学才是唯一能完全理解这种紧密联系的理论"。①

由于语言符号具有本书中所说的这些特性，因此，在人文科学各部门中，语言学是比较容易使用数学方法的。美国语言学家萨丕尔（E. Sapir）早就说过："印欧比较语言学的许多公式，其精密程度和其规律性令人想起自然科学的公式或者叫做定律"。②但是，作为人文科学的语言学当然也应该具有一般人文科学的特性，语言在本质上是一种社会现象，在语言学研究中，必不可免地要遵从人文科学的一般方法论原则，从这个意义上说，数学方法在语言学中的应用又是有条件的和有限度的。恩格斯曾经指出："把化学过程无条件地归结为纯粹的机械过程，是把研究的领域，至少是把化学的研究领域不适当地缩小了"。③作为自然现象的化学过程尚且不能无条件地归结为纯粹的机械过程，那么，作为社会现象的语言就更应该是这样了，因此我们在语言研究中，一定要从具体对象的"质"的特点出发，结合着使用对口径的数学方法，这样，才能使数学方法起到恰到好处的作用。数学和语言学是人类古老文明的两极，它们的"远缘杂交"，一定会开出美丽的花朵，结出丰硕的果实来。

---

① *J. Whatmough, Mathematical linguistics*，载《第八届国际语言学家大会文集》，第1卷，第218页，奥斯陆，1957年。

② 萨丕尔，《作为一门科学的语言学的地位》，(*Language*)，*Vol*, 5. 1929年，第214页。

③ 恩格斯，自然辩证法，人民出版社、1960年，第207页。